J. Yamato et al.
8/21/03
Q76950
1 of 1

日 本 国 特 許 庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出願年月日
Date of Application: ２００２年 ８月２８日

出願番号
Application Number: 特願２００２－２４９０４９

[ST.10/C]: [ＪＰ２００２－２４９０４９]

出願人
Applicant(s): 日本電気株式会社

２００３年 ５月２０日

特許庁長官
Commissioner,
Japan Patent Office 太田信一郎

出証番号 出証特２００３－３０３７４２５

【書類名】　特許願
【整理番号】　35001164
【提出日】　平成14年 8月28日
【あて先】　特許庁長官　殿
【国際特許分類】　G06F 13/00
【発明者】
　【住所又は居所】　東京都港区芝五丁目７番１号　日本電気株式会社内
　【氏名】　大和　純一
【発明者】
　【住所又は居所】　東京都港区芝五丁目７番１号　日本電気株式会社内
　【氏名】　菊地　芳秀
【発明者】
　【住所又は居所】　東京都港区芝五丁目７番１号　日本電気株式会社内
　【氏名】　金子　裕治
【特許出願人】
　【識別番号】　000004237
　【氏名又は名称】　日本電気株式会社
【代理人】
　【識別番号】　100103090
　【弁理士】
　【氏名又は名称】　岩壁　冬樹
　【電話番号】　03-3811-3561
【選任した代理人】
　【識別番号】　100114720
　【弁理士】
　【氏名又は名称】　須藤　浩
　【電話番号】　03-3811-3561
【手数料の表示】
　【予納台帳番号】　050496

【納付金額】　21,000円

【その他】　国等の委託研究の成果に係る特許出願（平成１３年度新エネルギー・産業技術総合開発機構「基盤技術研究促進事業（民間基盤技術研究支援制度）「大規模・高信頼サーバの研究」」に関する委託研究、産業活力再生特別措置法第３０条の適用を受けるもの）

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【包括委任状番号】　0102926

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　データ複製システム、中継装置、データ送受信方法およびストレージ内のデータを複製するためのプログラム

【特許請求の範囲】

【請求項１】　第一のストレージ内のデータを通信ネットワークを介して第二のストレージにミラーリングまたはバックアップするデータ複製システムにおいて、

前記第一のストレージから前記第二のストレージに転送されるデータを中継する中継装置であって、前記第一のストレージが災害によって稼働できない状態になっても、稼働を継続できるとあらかじめ算定された位置に設置された中継装置を備え、

前記第一のストレージは、データ転送の制御を行うデータ転送処理手段を含み、

前記データ転送処理手段は、前記中継装置に対してデータ転送を完了したときに、前記第二のストレージに対してデータ転送を完了したとみなす

ことを特徴とするデータ複製システム。

【請求項２】　中継装置は、第一のストレージから受信したコマンドおよびデータを記憶するための不揮発性記憶手段と、データの中継制御を行う中継処理手段とを含み、

前記中継処理手段は、第一のストレージから受信したコマンドおよびデータを前記不揮発性記憶手段に記憶し、記憶されたコマンドおよびデータを任意のタイミングで第二のストレージに送信する

請求項１記載のデータ複製システム。

【請求項３】　複数の中継装置が設けられ、

第一のストレージにおけるデータ転送処理手段は、第一のストレージ内のデータを同時に複数の中継装置に向けて送信する

請求項１または請求項２に記載のデータ複製システム。

【請求項４】　第一のストレージから第二のストレージに転送されるデータを中継する中継装置であって、

前記第一のストレージから受信したデータを記憶するための記憶手段と、データの中継制御を行う中継処理手段とを含み、

前記中継処理手段は、前記第一のストレージから受信したデータを記憶手段に記憶し、記憶手段にデータを記憶したときに前記第一のストレージに対する応答を送信し、前記第二のストレージに対して記憶手段に記憶したデータを送信することを特徴とする中継装置。

【請求項５】　データを送信する送信元からデータを受信する送信先にデータを送信するデータ送受信方法において、

前記送信元では、送信される元データから少なくとも１つのエラー訂正のための冗長データを作成し、元データと冗長データとを別々のデータ送信単位で送信する

ことを特徴とするデータ送受信方法。

【請求項６】　送信先では、元データと冗長データとの集合であるデータ群のすべてについて受信を完了する前に、元データについて部分的にエラー訂正処理を実行できる前記データ群の一部を受信した段階で、エラー訂正処理を実行する

請求項５記載のデータ送受信方法。

【請求項７】　送信元では、元データを分割データに分割し、それらの分割データのうちの１つまたは複数が消失しても元データを復元可能な冗長データを作成する

請求項５または請求項６に記載のデータ送受信方法。

【請求項８】　冗長データとしてパリティデータまたはＥＣＣを用いる請求項７記載のデータ送受信方法。

【請求項９】　冗長データとして送信データの複製データを用いる

請求項５から請求項７のうちのいずれか１項に記載のデータ送受信方法。

【請求項１０】　元データと冗長データとを、別々の通信ネットワークに送出する

請求項５から請求項９のうちのいずれか１項に記載のデータ送受信方法。

【請求項１１】　第一のストレージ内のデータを第二のストレージに通信ネットワークを介してミラーリングまたはバックアップするデータ複製システムにお

いて、

前記第一のストレージは、データ転送の制御を行うデータ転送処理手段と、送信される元データから少なくとも１つのエラー訂正のための冗長データを作成する冗長化手段とを含み、

前記データ転送処理手段は、元データと前記冗長化手段が作成した冗長データとを別々のデータ送信単位で送信する

ことを特徴とするデータ複製システム。

【請求項１２】　第二のストレージは、第一のストレージから受信した冗長データを用いてエラー訂正処理を行うデータ復元手段と、データ復元手段が復元したデータを記憶媒体に格納する格納処理手段とを含み、

前記データ復元手段は、前記第一のストレージから元データと冗長データとの集合であるデータ群のすべてについて受信を完了する前に、元データについて部分的にエラー訂正処理を実行できる前記データ群の一部を受信した段階で、エラー訂正処理を実行する

請求項１１記載のデータ複製システム。

【請求項１３】　第一のストレージにおける冗長化手段は、元データを分割データに分割し、それらの分割データのうちの１つまたは複数が消失しても元データを復元可能な冗長データを作成する

請求項１１または請求項１２に記載のデータ複製システム。

【請求項１４】　冗長化手段は、冗長データとしてパリティデータまたはＥＣＣを用いる請求項１３記載のデータ複製システム。

【請求項１５】　冗長化手段は、冗長データとして、元データの複製データを作成する

請求項１１から請求項１３のうちのいずれか１項に記載のデータ複製システム。

【請求項１６】　データ転送処理手段は、元データと冗長データとを、別々の通信ネットワークに送出する

請求項１１から請求項１５のうちのいずれか１項に記載のデータ複製システム。

【請求項１７】　稼働系のストレージ内のデータを通信ネットワークを介して待機系のストレージにミラーリングするデータ複製システムにおいて、

前記稼働系のストレージは、データの書き込み要求が発生すると、書き込み対象のデータと遅延書き込み要求とを前記待機系のストレージに送信する遅延書き込み要求手段と、そのデータの状態であればアプリケーションがそのまま動作を再開可能な再開可能ポイントであることを知らせるための再開可能ポイント通知を上位装置から受けると遅延書き込み実行要求を前記待機系のストレージに送信する書き込み実行要求手段とを含み、

前記待機系のストレージは、データを一時的に記憶する一時記憶手段と、受信したデータを遅延書き込み要求に応じて前記一時記憶手段に格納するとともに、遅延書き込み実行要求を受信すると前記一時記憶手段に格納されているデータを記憶媒体に格納する格納処理手段とを含む

ことを特徴とするデータ複製システム。

【請求項１８】　稼働系のストレージにおける遅延書き込み要求手段と遅延書き込み実行要求手段とは、遅延書き込み要求と遅延書き込み実行要求とを待機系のストレージに非同期に送信し、

前記待機系のストレージにおける格納処理手段は、一の遅延書き込み実行要求を受信すると、一つ前の遅延書き込み実行要求と前記一の遅延書き込み実行要求との間に送信された遅延書き込み要求に対応するデータを全て一時記憶手段に格納し終え、前記一つ前の遅延書き込み実行要求以前に送信されたデータを記憶媒体に格納し終えたときに、一時記憶手段に格納されているデータを記憶媒体に格納する

請求項１７記載のデータ複製システム。

【請求項１９】　待機系のストレージにおける格納処理手段は、稼働系に異常が生ずると、一時記憶手段に格納されているデータを破棄する

請求項１７または請求項１８に記載のデータ複製システム。

【請求項２０】　稼働系のストレージ内のデータを通信ネットワークを介して待機系のストレージにミラーリングするデータ複製システムにおいて、

前記稼働系のストレージは、データの書き込み要求が発生すると、書き込み対

象のデータと書き込み要求とを前記待機系のストレージに送信する書き込み要求手段と、そのデータの状態であればアプリケーションがそのまま動作を再開可能な再開可能ポイントであることを知らせるための再開可能ポイント通知を上位装置から受けるとスナップショット作成要求を前記待機系のストレージに送信するスナップショット作成要求手段とを含み、

前記待機系のストレージは、書き込み要求を受信したら前記書き込み要求に対応するデータを書き込むべき領域を割り当てて記憶媒体に前記データを格納するとともに、前記記憶媒体へのデータ格納状況を示す格納情報を更新し、スナップショット作成要求を受信したらスナップショットを作成するスナップショット作成手段を含む

ことを特徴とするデータ複製システム。

【請求項２１】　書き込み要求手段は、待機系ストレージのスナップショット作成手段がスナップショットを作成している場合には、スナップショット作成手段がスナップショット作成を完了した後に、書き込み要求を待機系のストレージに送信する請求項２０に記載のデータ複製システム。

【請求項２２】　稼働系のストレージにおける書き込み要求手段とスナップショット作成要求手段とは、書き込み要求とスナップショット作成要求とを待機系のストレージに非同期に送信し、

前記待機系のストレージにおけるスナップショット作成手段は、各書き込み要求を受信したときに、受信した書き込み要求の直前のスナップショット作成要求に基づくスナップショットの作成を完了するまで、書き込み要求に対応するデータを記憶媒体に格納するのを待ち、

一のスナップショット作成要求を受信すると、一つ前のスナップショット作成要求に基づくスナップショットの作成を完了し、前記一つ前のスナップショット作成要求と前記一のスナップショットとの間に送信された書き込み要求に対応するデータを全て記録媒体に格納し終えたときに、スナップショットを作成する

請求項２０に記載のデータ複製システム。

【請求項２３】　待機系のストレージにおけるスナップショット作成手段は、稼働系に異常が生ずると、直前のスナップショット作成後にデータが格納された

記憶媒体の領域を未使用状態として解放し、格納情報を直前のスナップショット作成時の状態に戻す請求項２０ないし請求項２２のうちのいずれか１項に記載のデータ複製システム。

【請求項２４】　稼働系のストレージを使用する稼働系の上位装置は、再開可能ポイントになると前記稼働系のストレージに再開可能ポイント通知を送信する再開可能ポイント通知手段を含み、

待機系のストレージを使用する待機系の上位装置は、前記稼働系の異常を検出すると、異常の発生を待機系のストレージに通知して待機系のストレージの状態を前記再開可能ポイントに対応する状態にするように促し、待機系のストレージの状態が前記再開可能ポイントに対応する状態になったときに処理を再開する

請求項１９または請求項２３に記載のデータ複製システム。

【請求項２５】　稼働系のストレージを使用する稼働系の上位装置は、待機系のストレージを使用する待機系の上位装置に、稼働系の処理実行状態を示す実行イメージを転送する実行イメージ転送手段と、

実行イメージ転送手段が実行イメージを転送するタイミングで、稼働系のストレージに再開可能ポイント通知を送信する任意時点再開可能ポイント通知手段とを含み、

待機系のストレージを使用する待機系の上位装置は、前記実行イメージ転送手段が転送する実行イメージを保存する実行イメージ保存手段を含み、

実行イメージ転送手段は、任意の時点における実行イメージを待機系の上位装置に転送し、

待機系のストレージを使用する待機系の上位装置は、前記稼働系の異常を検出すると、異常の発生を待機系のストレージに通知して、待機系のストレージの状態を実行イメージの転送タイミングに対応する状態にするように促し、待機系のストレージの状態が実行イメージの転送タイミングに対応する状態になった場合に実行イメージを用いて処理を再開する

請求項１９または請求項２３に記載のデータ複製システム。

【請求項２６】　稼働系の上位装置における実行イメージ転送手段は、前回転送した実行イメージから変更された部分のみを転送する

請求項２５記載のデータ複製システム。

【請求項２７】　第一のストレージ内のデータを通信ネットワークおよび中継装置を介して第二のストレージにミラーリングまたはバックアップするデータ複製システムにおける前記中継装置内に設けられているコンピュータに、

前記第一のストレージから受信したデータを中継装置内の記憶媒体に記憶する処理と、

データを中継装置内の記憶媒体に記憶したときに前記第一のストレージに対する応答を送信する処理と、

中継装置内の記憶媒体に記憶したデータを前記第二のストレージに送信する処理とを

実行させるストレージ内のデータを複製するためのプログラム。

【請求項２８】　第一のストレージ内のデータを第二のストレージに通信ネットワークを介してミラーリングまたはバックアップするデータ複製システムにおける前記第一のストレージ内に設けられているコンピュータに、

送信される元データから少なくとも１つのエラー訂正のための冗長データを作成する処理と、

元データと冗長データとを別々のデータ送信単位で送信する処理とを実行させるストレージ内のデータを複製するためのプログラム。

【請求項２９】　稼働系のストレージ内のデータを通信ネットワークを介して待機系のストレージにミラーリングするデータ複製システムにおける前記稼働系のストレージ内に設けられているコンピュータに、

データの書き込み要求が発生すると、書き込み対象のデータと一時記憶装置にデータを格納することを指示する遅延書き込み要求とを前記待機系のストレージに送信する処理と、

そのデータの状態であればアプリケーションがそのまま動作を再開可能なポイントであることを知らせるための再開可能ポイント通知を上位装置から受けると、一時記憶装置に格納されているデータを記憶媒体に格納することを指示する遅延書き込み実行要求を前記待機系のストレージに送信する処理とを実行させるストレージ内のデータを複製するためのプログラム。

【請求項３０】　稼働系のストレージ内のデータを通信ネットワークを介して待機系のストレージにミラーリングするデータ複製システムにおける前記稼働系のストレージ内に設けられているコンピュータに、

データの書き込み要求が発生すると、書き込み対象のデータと記憶媒体にデータを書き込むことを指示する書き込み要求とを前記待機系のストレージに送信する処理と、そのデータの状態であればアプリケーションがそのまま動作を再開可能なポイントであることを知らせるための再開可能ポイント通知を上位装置から受けるとスナップショット作成要求を前記待機系のストレージに送信する処理を実行させるストレージ内のデータを複製するためのプログラム。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、データ複製システム、中継装置、データ送受信方法およびストレージ内のデータを複製するためのプログラムに関する。

【０００２】

【従来の技術】

災害等が発生してもコンピュータシステムの機能を維持できるようにするため、正常系（稼働系）のシステムと待機系のシステムとが設けられたコンピュータシステムが実現されている。例えば、ＥＭＣ　Ｃｏｒｐｏｒａｔｉｏｎ（イーエムシー　コーポレーション）は、正常系のストレージと待機系のストレージとを用いてミラーリングを行うシステムを実現している。このシステムに関する情報は、「http://www.emc2.co.jp/local/ja/JP/products/product_pdfs/srdf/srdf.pdf」というＵＲＬで公開されている。

【０００３】

また、特開２０００－３０５８５６公報には、メインセンター（正常系のシステム）とリモートセンター（待機系のシステム）とでデータの二重化を図るシステムが記載されている。

【０００４】

一般に、正常系のシステムでは、正常系のストレージとそのストレージを使用

する正常系のホストコンピュータ（以下、ホストという。）とが接続される。待機系のシステムも同様である。そして、正常系のストレージと待機系のストレージとが、例えば、専用回線やインターネット等の通信ネットワークを介して接続される。

【０００５】

「http://www.emc2.co.jp/local/ja/JP/products/product_pdfs/srdf/srdf.pdf」において公開されているシステムや、特開２０００－３０５８５６公報に記載されたシステムでは、正常系のシステムから待機系のシステムに直接データを送信している。しかし、正常系のシステムから中継装置にデータを送信し、中継装置から待機系のシステムにデータを送信することもある。一般に、中継装置は、正常系のシステムから受信したデータを待機系のシステムに送信し、待機系のシステムから受信完了の通知を受けたときに、正常系のシステムに待機系へのデータ転送が完了したことを通知している。そして、正常系のシステムは、中継装置にデータを送信した場合、データ転送が完了した旨の通知を中継装置から受信してから次の処理を開始する。

【０００６】

また、一般に、通信ネットワークを介してデータを送受信する場合、データを分割してデータ送信単位毎に送受信する。このデータ送信単位は、通信プロトコル毎に異なる。ここでは、パケットをデータ送信単位とする場合を例に説明する。パケット等のデータ送信単位には、送信対象のデータだけでなく、送信過程でのデータの誤り（データ化け）を検出するためのエラー検出コードも含まれる。エラー検出コードとしては、チェックサムデータやＣＲＣ（Cyclic Redundancy Check）データ等がある。データを受信した装置は、エラー検出コードによってデータの誤りを検出すると、そのパケットを廃棄する。

【０００７】

また、データに誤りが生じていない場合でも、通信ネットワークに輻輳（通信負荷が高い状態）が生じると、パケットは通信ネットワーク上で廃棄される。パケットが廃棄され、送信先からの応答が得られない場合、送信元は再度パケットを送信する。なお、通信ネットワークに輻輳が生じたときにパケットの廃棄を開

始すると、送信元による再送信によってさらに通信負荷が高くなったり、通信ネットワークの利用率が低くなったりすることがある。このような問題を回避するため、近年、通信負荷が所定のしきい値を越えたときに通信ネットワークを構成する機器が任意にパケットを廃棄する方式が採用されている。この方式は、ＲＥＤ（Random Early Detection）と呼ばれている。

【０００８】

また、データを受信した装置がデータの誤りを訂正することができるデータ伝送方式も提案されている。例えば、特開昭５７－１３８２３７号公報には、送信対象のデータと、そのデータから作成されたパリティビットとを別々に送信し、送信対象のデータとパリティビットとを受信した装置が誤り訂正を行うデータ伝送方式が記載されている。

【０００９】

なお、以下、ミラーリングとバックアップとを以下のように区別するものとする。「ミラーリング」とは、あるストレージに対してホストから書込コマンド（write コマンド）が出力されたことを契機として、そのストレージを含む２つ以上のストレージに対して同じデータを書き込むことと定義する。レプリケーションもミラーリングの一種として扱う。また、「バックアップ」とは、ストレージに対するwrite コマンドを契機とせず、任意のタイミングで、あるストレージの内容を他のストレージに複写することと定義する。

【００１０】

また、コンピュータシステムにおいて、実行中のプログラムを任意の時点で中断させ、後に再開させる技術が実現されている。例えば、ＳＸシリーズという名称で販売されている日本電気株式会社製のスーパコンピュータでは、任意の時点におけるプロセスの実行状態（例えば、メモリやレジスタの状態等）を保存して、プログラムを中断・再開するようにしている。この機能は、チェックポイント・リスタート機能と呼ばれる場合もある。

【００１１】

【発明が解決しようとする課題】

正常系のシステムから中継装置を介して待機系のシステムにデータを送信する

場合には、以下のような問題があった。中継装置は、正常系のシステムから受信したデータを待機系のシステムに送信し、待機系のシステムから受信完了の通知を受けたときに、転送が完了したことを正常系のシステムに通知する。そして、正常系のシステムは、転送が完了した旨の通知を中継装置から受信しなければ、次の処理に移行しない。そのため、正常系のシステムが中継装置にデータを送信してから、次の処理を開始できるまでの時間がかかってしまうという問題があった。特に、災害対策として中継装置や待機系のシステムを設ける場合には、正常系のシステムの遠隔地に中継装置を配置し、待機系のシステムをより遠い場所に配置する。その結果、正常系システムと中継装置との間でデータや通知を送受信する時間や、中継装置と待機系システムとの間でデータや通知を送受信する時間がかかってしまい、正常系のシステムにおける次の処理の開始が遅れてしまう。

【００１２】

例えば１００ｋｍ離れた地点との通信におけるラウンドトリップ時間（データを送信してから応答を得るまでの往復時間）は、ｍｓ（ミリ秒）のオーダである。ホストは、μｓ（マイクロ秒）のオーダで各処理を進める。従って、送信に対する応答を得るまでの時間は、処理の遅れの原因となる。

【００１３】

また、データの送信過程において、パケット等のデータ送信単位で廃棄されると、送信元となる正常系のストレージはデータを再送信することになる。すると、データ送信完了までに一層時間がかかり、次の処理の開始がさらに遅れてしまう。

【００１４】

また、正常系のストレージから待機系のストレージにデータのバックアップ処理を行うときには、データの送信距離が長いことに加えて、送信すべきデータの量が多くなるため、送信時間が一層増大してしまう。

【００１５】

また、正常系のシステムに異常が生じたときに、待機系のシステムですぐに処理を開始できるとは限らなかった。例えば、ある処理Ｘを開始する際には、ストレージに対するデータＡおよびデータＢの書き込みが完了していなければならな

いとする。正常系のホストがデータＡを書き込むwrite コマンドをストレージに出力してミラーリングを行うと、待機系のストレージにもデータＡは反映される。その後、正常系のシステムに異常が発生したとする。この場合、待機系のストレージには、データＢが書き込まれていないため、すぐに処理Ｘを開始できない。待機系のシステムで処理を開始する場合には、待機系のストレージにデータＢを書き込んで処理Ｘを開始するか、あるいは、データＡを削除して処理Ｘの前の処理からやり直す必要がある。そのため、待機系システムでの処理の再開に時間がかかってしまっていた。

【００１６】

また、所定のデータ記録状態になっていれば処理を再開できるような機能がアプリケーションプログラムによって実現されていない場合がある。例えば、上記の例においてホストがストレージにＡ，Ｂを書き込んだならば処理Ｘから処理を再開できるようにした機能が実現されていない場合がある。以下、ストレージが所定のデータ記録状態になっていれば処理を再開できるような機能を「再開機能」と記す。アプリケーションプログラムによって再開機能が実現されていない場合においても、正常系で異常が生じたならば、待機系で迅速に処理を再開できることが好ましい。

【００１７】

そこで本発明は、中継装置を介して待機系のシステムにデータを送信する場合に、正常系のシステムが早く次の処理を開始できるようにすることを目的とする。また、送信過程でデータが廃棄されたとしても迅速にデータの送受信を完了できるようにすることを目的とする。また、正常系のシステムで異常が発生したときに、すぐに待機系のシステムが処理を進められるようにすることを目的とする。

【００１８】

【課題を解決するための手段】

本発明によるデータ複製システムは、第一のストレージ内のデータを通信ネットワークを介して第二のストレージにミラーリングまたはバックアップするデータ複製システムにおいて、第一のストレージから第二のストレージに転送される

データを中継する中継装置であって、第一のストレージが災害によって稼働できない状態になっても、稼働を継続できるとあらかじめ算定された位置に設置された中継装置を備え、第一のストレージは、データ転送の制御を行うデータ転送処理手段を含み、データ転送処理手段は、中継装置に対してデータ転送を完了したときに、第二のストレージに対してデータ転送を完了したとみなすことを特徴とする。

【００１９】

中継装置は、第一のストレージから受信したコマンドおよびデータを記憶するための不揮発性記憶手段と、データの中継制御を行う中継処理手段とを含み、中継処理手段は、第一のストレージから受信したコマンドおよびデータを不揮発性記憶手段に記憶し、記憶されたコマンドおよびデータを任意のタイミングで第二のストレージに送信する構成であってもよい。そのような構成によれば、中継装置と第二のストレージとが接続される通信ネットワークの運用コストを低減することができる。

【００２０】

複数の中継装置が設けられ、第一のストレージにおけるデータ転送処理手段は、第一のストレージ内のデータを同時に複数の中継装置に向けて送信してもよい。そのような構成によれば、データの転送をより迅速に行うことができる。

【００２１】

また、本発明による中継装置は、第一のストレージから第二のストレージに転送されるデータを中継する中継装置であって、第一のストレージから受信したデータを記憶するための記憶手段と、データの中継制御を行う中継処理手段とを含み、中継処理手段は、第一のストレージから受信したデータを記憶手段に記憶し、記憶手段にデータを記憶したときに第一のストレージに対する応答を送信し、第二のストレージに対して記憶手段に記憶したデータを送信することを特徴とする。

【００２２】

また、本発明によるデータを送信する送信元からデータを受信する送信先にデータを送信するデータ送受信方法において、送信元では、送信される元データか

ら少なくとも１つのエラー訂正のための冗長データを作成し、元データと冗長データとを別々のデータ送信単位で送信することを特徴とする。

【００２３】

送信先では、元データと冗長データとの集合であるデータ群のすべてについて受信を完了する前に、元データについて部分的にエラー訂正処理を実行できるデータ群の一部を受信した段階で、エラー訂正処理を実行することが好ましい。そのような方法によれば、送信過程でデータの一部が廃棄されても、送信元はデータを再度送信する必要がない。

【００２４】

例えば、送信元では、元データを分割データに分割し、それらの分割データのうちの１つまたは複数が消失しても元データを復元可能な冗長データを作成する。

【００２５】

例えば、冗長データとしてパリティデータまたはＥＣＣ（Error Correcting Code）を用いればよい。

【００２６】

また、冗長データとして送信データの複製データを用いてもよい。

【００２７】

元データと冗長データとを、別々の通信ネットワークに送出してもよい。そのような方法によれば、一方の通信ネットワークで障害等が発生しても、もう一方の通信ネットワークから受信したデータによって処理を進めることができる。

【００２８】

また、本発明によるデータ複製システムは、第一のストレージ内のデータを第二のストレージに通信ネットワークを介してミラーリングまたはバックアップするデータ複製システムにおいて、第一のストレージは、データ転送の制御を行うデータ転送処理手段と、送信される元データから少なくとも１つのエラー訂正のための冗長データを作成する冗長化手段とを含み、データ転送処理手段は、元データと冗長化手段が作成した冗長データとを別々のデータ送信単位で送信することを特徴とする。

【００２９】

第二のストレージは、第一のストレージから受信した冗長データを用いてエラー訂正処理を行うデータ復元手段と、データ復元手段が復元したデータを記憶媒体に格納する格納処理手段とを含み、データ復元手段は、第一のストレージから元データと冗長データとの集合であるデータ群のすべてについて受信を完了する前に、元データについて部分的にエラー訂正処理を実行できるデータ群の一部を受信した段階で、エラー訂正処理を実行することが好ましい。そのような構成によれば、送信過程でデータの一部が廃棄されても、送信元はデータを再度送信する必要がない。

【００３０】

例えば、第一のストレージにおける冗長化手段は、元データを分割データに分割し、それらの分割データのうちの１つまたは複数が消失しても元データを復元可能な冗長データを作成する。

【００３１】

冗長化手段は、冗長データとしてパリティデータまたはＥＣＣを用いてもよい。

【００３２】

冗長化手段は、冗長データとして、元データの複製データを作成してもよい。

【００３３】

データ転送処理手段は、元データと冗長データとを、別々の通信ネットワークに送出してもよい。そのような構成によれば、一方の通信ネットワークで障害等が発生しても、もう一方の通信ネットワークから受信したデータによって処理を進めることができる。

【００３４】

また、本発明によるデータ複製システムは、稼働系のストレージ内のデータを通信ネットワークを介して待機系のストレージにミラーリングするデータ複製システムにおいて、稼働系のストレージは、データの書き込み要求が発生すると、書き込み対象のデータと遅延書き込み要求とを待機系のストレージに送信する遅延書き込み要求手段と、そのデータの状態であればアプリケーションがそのまま

動作を再開可能な再開可能ポイントであることを知らせるための再開可能ポイント通知を上位装置から受けると遅延書き込み実行要求を待機系のストレージに送信する書き込み実行要求手段とを含み、待機系のストレージは、データを一時的に記憶する一時記憶手段と、受信したデータを遅延書き込み要求に応じて一時記憶手段に格納するとともに、遅延書き込み実行要求を受信すると一時記憶手段に格納されているデータを記憶媒体に格納する格納処理手段とを含むことを特徴とする。

【００３５】

稼働系のストレージにおける遅延書き込み要求手段と遅延書き込み実行要求手段とは、遅延書き込み要求と遅延書き込み実行要求とを待機系のストレージに非同期に送信し、待機系のストレージにおける格納処理手段は、一の遅延書き込み実行要求を受信すると、一つ前の遅延書き込み実行要求と一の遅延書き込み実行要求との間に送信された遅延書き込み要求に対応するデータを全て一時記憶手段に格納し終え、一つ前の遅延書き込み実行要求以前に送信されたデータを記憶媒体に格納し終えたときに、一時記憶手段に格納されているデータを記憶媒体に格納する。

【００３６】

待機系のストレージにおける格納処理手段は、稼働系に異常が生ずると、一時記憶手段に格納されているデータを破棄する。

【００３７】

また、本発明によるデータ複製システムは、稼働系のストレージ内のデータを通信ネットワークを介して待機系のストレージにミラーリングするデータ複製システムにおいて、稼働系のストレージは、データの書き込み要求が発生すると、書き込み対象のデータと書き込み要求とを待機系のストレージに送信する書き込み要求手段と、そのデータの状態であればアプリケーションがそのまま動作を再開可能な再開可能ポイントであることを知らせるための再開可能ポイント通知を上位装置から受けるとスナップショット作成要求を待機系のストレージに送信するスナップショット作成要求手段とを含み、待機系のストレージは、書き込み要求を受信したら書き込み要求に対応するデータを書き込むべき領域を割り当てて

記憶媒体にデータを格納するとともに、記憶媒体へのデータ格納状況を示す格納情報を更新し、スナップショット作成要求を受信したらスナップショットを作成するスナップショット作成手段を含むことを特徴とする。

【００３８】

例えば、書き込み要求手段は、待機系ストレージのスナップショット作成手段がスナップショットを作成している場合には、スナップショット作成手段がスナップショット作成を完了した後に、書き込み要求を待機系のストレージに送信する。

【００３９】

また、稼働系のストレージにおける書き込み要求手段とスナップショット作成要求手段とは、書き込み要求とスナップショット作成要求とを待機系のストレージに非同期に送信し、待機系のストレージにおけるスナップショット作成手段は、各書き込み要求を受信したときに、受信した書き込み要求の直前のスナップショット作成要求に基づくスナップショットの作成を完了するまで、書き込み要求に対応するデータを記憶媒体に格納するのを待ち、一のスナップショット作成要求を受信すると、一つ前のスナップショット作成要求に基づくスナップショットの作成を完了し、一つ前のスナップショット作成要求と一のスナップショットとの間に送信された書き込み要求に対応するデータを全て記録媒体に格納し終えたときに、スナップショットを作成してもよい。

【００４０】

待機系のストレージにおけるスナップショット作成手段は、稼働系に異常が生ずると、直前のスナップショット作成後にデータが格納された記憶媒体の領域を未使用状態として解放し、格納情報を直前のスナップショット作成時の状態に戻す。

【００４１】

また、稼働系のストレージを使用する稼働系の上位装置は、再開可能ポイントになると稼働系のストレージに再開可能ポイント通知を送信する再開可能ポイント通知手段を含み、待機系のストレージを使用する待機系の上位装置は、稼働系の異常を検出すると、異常の発生を待機系のストレージに通知して待機系のスト

レージの状態を再開可能ポイントに対応する状態にするように促し、待機系のストレージの状態が再開可能ポイントに対応する状態になったときに処理を再開する構成であってもよい。

【００４２】

また、稼働系のストレージを使用する稼働系の上位装置は、待機系のストレージを使用する待機系の上位装置に、稼働系の処理実行状態を示す実行イメージを転送する実行イメージ転送手段と、実行イメージ転送手段が実行イメージを転送するタイミングで、稼働系のストレージに再開可能ポイント通知を送信する任意時点再開可能ポイント通知手段とを含み、待機系のストレージを使用する待機系の上位装置は、実行イメージ転送手段が転送する実行イメージを保存する実行イメージ保存手段を含み、実行イメージ転送手段は、任意の時点における実行イメージを待機系の上位装置に転送し、待機系のストレージを使用する待機系の上位装置は、稼働系の異常を検出すると、異常の発生を待機系のストレージに通知して、待機系のストレージの状態を実行イメージの転送タイミングに対応する状態にするように促し、待機系のストレージの状態が実行イメージの転送タイミングに対応する状態になった場合に実行イメージを用いて処理を再開する構成であってもよい。

【００４３】

稼働系の上位装置における実行イメージ転送手段は、前回転送した実行イメージから変更された部分のみを転送してもよい。

【００４４】

また、本発明によるストレージ内のデータを複製するためのプログラムは、第一のストレージ内のデータを通信ネットワークおよび中継装置を介して第二のストレージにミラーリングまたはバックアップするデータ複製システムにおける中継装置内に設けられているコンピュータに、第一のストレージから受信したデータを中継装置内の記憶媒体に記憶する処理と、データを中継装置内の記憶媒体に記憶したときに第一のストレージに対する応答を送信する処理と、中継装置内の記憶媒体に記憶したデータを第二のストレージに送信する処理とを実行させる。

【００４５】

また、本発明によるストレージ内のデータを複製するためのプログラムは、第一のストレージ内のデータを第二のストレージに通信ネットワークを介してミラーリングまたはバックアップするデータ複製システムにおける第一のストレージ内に設けられているコンピュータに、送信される元データから少なくとも１つのエラー訂正のための冗長データを作成する処理と、元データと冗長データとを別々のデータ送信単位で送信する処理とを実行させる。

【００４６】

また、本発明によるストレージ内のデータを複製するためのプログラムは、稼働系のストレージ内のデータを通信ネットワークを介して待機系のストレージにミラーリングするデータ複製システムにおける稼働系のストレージ内に設けられているコンピュータに、

データの書き込み要求が発生すると、書き込み対象のデータと一時記憶装置にデータを格納することを指示する遅延書き込み要求とを待機系のストレージに送信する処理と、そのデータの状態であればアプリケーションがそのまま動作を再開可能なポイントであることを知らせるための再開可能ポイント通知を上位装置から受けると、一時記憶装置に格納されているデータを記憶媒体に格納することを指示する遅延書き込み実行要求を待機系のストレージに送信する処理とを実行させる。

【００４７】

また、本発明によるストレージ内のデータを複製するためのプログラムは、稼働系のストレージ内のデータを通信ネットワークを介して待機系のストレージにミラーリングするデータ複製システムにおける稼働系のストレージ内に設けられているコンピュータに、

データの書き込み要求が発生すると、書き込み対象のデータと記憶媒体にデータを書き込むことを指示する書き込み要求とを待機系のストレージに送信する処理と、そのデータの状態であればアプリケーションがそのまま動作を再開可能なポイントであることを知らせるための再開可能ポイント通知を上位装置から受けるとスナップショット作成要求を待機系のストレージに送信する処理を実行させる。

【００４８】

【発明の実施の形態】

以下、本発明の実施の形態を図面を参照して説明する。

【００４９】

実施の形態１．

図１は、本発明によるデータ複製システムの第１の実施の形態を示すブロック図である。図１に示すデータ複製システムにおいて、ストレージ１１が、ストレージ１１を使用するホスト（ホストコンピュータ）１０とローカルに接続されている。ストレージ１１は、インターネットや専用線等の通信ネットワーク（以下、ネットワークという。）１３を介して中継装置１５に接続されている。また、中継装置１５は、ネットワーク１４を介してストレージ１２に接続されている。

【００５０】

ストレージ１１，１２は、例えば、単体の磁気ディスク装置、光ディスク装置または光磁気ディスク装置である。ストレージ１１，１２として、単体の磁気ディスク装置、光ディスク装置または光磁気ディスク装置の集合であるディスクアレイ装置を使用することもできる。ホスト１０とストレージ１１とは、ＳＣＳＩ、ファイバチャネル（Fibre channel ）、イーサネット（登録商標）等で接続される。

【００５１】

図２は、図１に示すストレージ１１の構成例を示すブロック図である。なお、ストレージ１２の構成も、図２に示すような構成である。図２に示すように、ストレージ１１は、ストレージコントローラ１００とストレージ本体である記憶媒体１０１とを含む。ストレージコントローラ１００は、ホスト１０および他のストレージまたは中継装置と通信を行う通信部１０２、各処理のシーケンスを管理する処理シーケンサ１０３、記憶媒体１０１に対する処理命令の順序制御を行うＩＯスケジューラ１０４、ＩＯスケジューラ１０４が発行する処理命令に従って記憶媒体１０１の動作を制御する媒体処理部１０５、およびホスト１０等から記憶媒体１０１へのデータおよび記憶媒体１０１からホスト１０等へのデータを一時記憶するバッファメモリ１０６を含む。処理シーケンサ１０３は、例えば、プ

ログラムに従って動作するＣＰＵで実現される。

【００５２】

図２では、媒体制御部１０５と記憶媒体１０１との組み合わせを一組だけ含む場合を示した。ストレージ１１、１２が複数の記録媒体１０１を含み、それぞれの記憶媒体に対応して媒体制御部１０５が設けられていてもよい。ただし、個々のストレージ１１，１２に含まれるＩＯスケジューラ１０４は一つである。記憶媒体１０１が複数存在する場合、処理シーケンサ１０３は処理命令の対象となる記憶媒体を指定し、ＩＯスケジューラ１０４は、指定された記憶媒体に対応する媒体制御部１０５に処理を行わせる。

【００５３】

図３は、図１に示す中継装置１５の構成例を示すブロック図である。中継装置１５は、ストレージ１１，１２と通信を行う通信部１５０、中継処理のシーケンス管理を行う中継処理部１５１、およびストレージ１１，１２から受信したデータを一時記憶するバッファメモリ１５２を含む。中継処理部１５１は、例えば、プログラムに従って動作するＣＰＵで実現される。

【００５４】

図１および図２に示すように、ストレージ１１には通信部１０２が設けられ、ストレージ１１自身が主体的にデータを転送する。すなわち、ホスト１０がストレージ１１からデータを読み出して中継装置１５にデータを送信するのではなく、ストレージ１１が直接中継装置１５にデータを送信する。ストレージ１２も、中継装置１５から直接データを受信する。ただし、ストレージ１１が中継装置１５にデータを送信するタイミングは、ホスト１０が指示する。

【００５５】

また、中継装置１５は、ストレージ１１，１２の設置場所において地震等の災害が発生したときに、災害の影響が波及すると想定される範囲の外に設置される。すなわち、ストレージ１１，１２が災害によって稼働できない状態になっても、その位置であれば稼働を継続できる位置（ストレージ１１，１２からの距離）があらかじめ算定され、中継装置１５は、算定された位置に設置されている。例えば、ストレージ１１があらかじめ想定した震度（例えば震度６～７）の地震に

被災したときに、中継装置１５には破損等が生じない程度の震度になるような位置に、中継装置１５が設置される。さらに、ストレージ１１またはストレージ１２と中継装置１５との間のデータ転送時間が、ストレージ１１とストレージ１２との間で直接データを転送した場合のデータ転送時間よりも短くなるように、中継装置１５が設置される。よって、データ複製システムは、耐災害データ管理システムとして機能する。

【００５６】

次に、データ複製システムの動作を、図４および図５のフローチャートを参照して説明する。ここでは、バックアップ処理として、データの転送元としてのストレージ１１が、データの転送先としてのストレージ１２に向けてデータを転送する場合を例にする。図４は、ストレージコントローラ１００における処理シーケンサ１０３の動作を示すフローチャートであり、図５は中継装置１５における中継処理部１５１の動作を示すフローチャートである。

【００５７】

ホスト１０は、バックアップ処理を実行するときに、ストレージ１１のストレージコントローラ１００における通信部１０２に対して複写コマンドを出力する。複写コマンドは、バックアップを指示するコマンドである。複写コマンドには、複写対象のデータの範囲すなわちバックアップ処理対象のデータ範囲を指定する情報および複写先のストレージを指定する情報が含まれている。通信部１０２は、複写コマンドを受け取ると、複写コマンドを処理シーケンサ１０３に渡し、複写処理を開始することを処理シーケンサ１０３に指示する。

【００５８】

処理シーケンサ１０３は、複写コマンドを受け取ると、図４に示すように、まず、複写対象のデータの範囲の先頭のブロックをデータ転送対象のブロックとして設定する（ステップＳ１００）。次いで、通信部１０２に、複写コマンドで指定された複写先のストレージに対応した中継装置に対してwrite コマンドを送信するように指示する（ステップＳ１０１）。write コマンドには、ブロックサイズすなわちデータ量を示す情報が含まれている。また、データ転送対象のデータの読み出し要求をＩＯスケジューラ１０４に登録する。ＩＯスケジューラ１０４

は、読み出し要求に応じて、媒体制御部１０５に、データ転送対象のデータの読み出し指示を行う。媒体制御部１０５は、読み出し指示に従って、データ転送対象のデータを記憶媒体１０１からバッファメモリ１０６に出力させる（ステップＳ１０２）。媒体制御部１０５は、データ転送対象の全てのデータが記憶媒体１０１からバッファメモリ１０６に出力されると、読み出し完了通知を処理シーケンサ１０３に出力する。

【００５９】

そして、処理シーケンサ１０３は、通信部１０２を介して入力される中継装置１５からの受信準備完了のメッセージと、媒体制御部１０５からの読み出し完了通知との双方を待ち（ステップＳ１０３）。双方が入力されたら、バッファメモリ１０６に記憶されたデータを中継装置１５に転送するように通信部１０２に指示する。通信部１０２は、指示に応じてバッファメモリ１０６に記憶されたデータを中継装置１５に送信する（ステップＳ１０４）。そして、処理シーケンサ１０３は、通信部１０２を介して入力される中継装置１５からの受信完了のメッセージを待つ（ステップＳ１０５）。

【００６０】

通信部１０２は、バッファメモリ１０６に記憶された全てのデータを中継装置１５に送信し、中継装置１５から受信完了のメッセージを受けたら、受信完了のメッセージを処理シーケンサ１０３を出力する。処理シーケンサ１０３は、受信完了のメッセージを入力すると、複写対象の全てのデータの中継装置１５への転送が完了したか否か確認する（ステップＳ１０６）。完了していなければ、複写対象のデータの範囲の次のブロックをデータ転送対象のブロックとして設定し（ステップＳ１０８）、ステップＳ１０１に戻る。

【００６１】

複写対象の全てのデータの転送が完了していれば、処理シーケンサ１０３は、ホスト１０に対して完了通知を出力するように通信部１０２に指示し（ステップＳ１０７）、処理を終了する。

【００６２】

なお、１ブロックのデータ量は、あらかじめシステムに設定されている量であ

る。また、処理シーケンサ１０３は、１ブロックのデータ量を、転送先のストレージ１２のバッファメモリ１０６および中継装置１５のバッファメモリ１５２の容量に応じて変化させるようにしてもよい。

【００６３】

処理シーケンサ１０３がＩＯスケジューラ１０４に登録するものとして、処理の種類（読み出し／書き込み）と、処理を識別するためのＩＤと、処理の対象となる記憶媒体１０１中の領域を示す情報と、処理の対象となるバッファメモリ１０６の領域を示す情報とがある。なお、処理シーケンサ１０３が複数の読み出し処理命令や複数の書き込み処理命令を指示する場合もあり、処理を識別するためのＩＤは、このような場合に各処理を識別するために用いられる。従って、図４に示された処理では、処理シーケンサ１０３は、データ転送対象のデータの読み出し要求を登録する際に、処理の種類、読み出し処理を識別するためのＩＤを登録するとともに、処理の対象となる記憶媒体１０１中の領域としてデータ転送対象のブロックを登録する。また、記憶媒体が複数ある場合には、処理シーケンサ１０３は、記憶媒体を特定するための情報も登録する。そして、処理シーケンサ１０３は、登録時に指定したＩＤによって、どの処理が完了したのかを判別する。

【００６４】

ＩＯスケジューラ１０４は、処理シーケンサ１０３によって登録された処理を記録する。そして、記録されている処理をあらかじめ決められているアルゴリズムに従って選択し、取り出した処理を媒体制御部１０５に実行させる。あらかじめ決められているアルゴリズムとして、例えば、登録された順がある。また、記憶媒体１０１の磁気ヘッドや光ヘッドの現在位置から処理対象の位置までの移動距離が最も小さくなる処理を最初に選択してもよい。また、ストレージ１１が複数の記録媒体１０１を含むディスクアレイ装置であってそれぞれの記憶媒体に対応して媒体制御部１０５が設けられている場合には、ＩＯスケジューラ１０４は、処理を取り出そうとした記憶媒体１０１に対応した媒体制御部１０５が担当する記憶媒体１０１を対象とした処理を選択する。

【００６５】

媒体制御部１０５の実行中の処理が完了すると、ＩＯスケジューラ１０４は次の処理を取り出す。処理が読み出し処理であった場合には、媒体制御部１０５は、記憶媒体１０１に、記憶媒体１０１における指定された領域からデータを読み出させて、読み出したデータをバッファメモリ１０６の指定された領域に書き込ませる。処理が書き込み処理であった場合には、媒体制御部１０５は、記憶媒体１０１に、バッファメモリ１０６の指定された領域のデータを、記憶媒体１０１における指定された領域に書き込ませる。そして、媒体制御部１０５は、処理が完了したときには、処理シーケンサ１０３に処理の完了を通知する。

【００６６】

通信部１０２は、外部から入力されたデータが、コマンド、受信準備完了のメッセージ、受信完了のメッセージ等の制御系メッセージであった場合には、入力されたデータを処理シーケンサ１０３に渡す。また、ストレージ１１に書き込まれるべきデータであった場合には、データを格納すべき場所を処理シーケンサ１０３に問い合わせる。そして、処理シーケンサ１０３から指定されたバッファメモリ１０６中の領域にデータを格納する。

【００６７】

また、通信部１０２は、処理シーケンサ１０３から指定されたコマンドまたは完了通知を、指定された中継装置またはホストに向けて送信する処理も行う。さらに、処理シーケンサ１０３から指定されたバッファメモリ１０６中のデータを、指定された中継装置またはホストに向けて送信する処理も行う。

【００６８】

なお、ホスト１０は、ストレージ１１から完了通知を受けたら、すなわち、中継装置１５へのデータ転送が完了したら、ストレージ１１からストレージ１２へのデータ転送が完了したと認識する。

【００６９】

次に、中継装置１５の動作を説明する。中継装置１５において、ストレージ１１，１２からのデータ（コマンドまたは転送されるデータ）は、通信部１５０で受信される。通信部１５０は、ストレージ１１がステップＳ１０１において送信したwrite コマンドを受信すると、write コマンドを中継処理部１５１に渡す。

【００７０】

中継処理部１５１は、図５に示すように、ストレージ１１から送られてくるデータを格納するのに必要な領域をバッファメモリ１５２に確保する（ステップＳ１２０）。中継処理部１５１は、write コマンドに含まれるデータ量の情報に基づいてデータの格納領域を確保する。そして、受信準備完了の通知をストレージ１１に送るように通信部１５０に指示する（ステップＳ１２１）。通信部１５０は、指示に応じて、ストレージ１１に受信準備完了のメッセージを送信する。また、write コマンドをストレージ１２に送るように通信部１５０に指示する（ステップＳ１２２）。通信部１５０は、指示に応じて、ストレージ１２にwrite コマンドを送信する。このwrite コマンドは、ストレージ１１から受信するデータをストレージ１２に書き込ませるためのwrite コマンドである。

【００７１】

次いで、ストレージ１１からデータが届くのを待ち（ステップＳ１２３）、データが届いて通信部１５０からデータを格納すべきバッファメモリ１５２の領域の問い合わせを受けると、ステップＳ１２０で確保した領域を通信部１５０に知らせる（ステップＳ１２４）。また、データのバッファメモリ１５２への格納の完了を待ち（ステップＳ１２５）、全てのデータがバッファメモリ１５２に格納されたことが通信部１５０から通知されると、ストレージ１１にwrite コマンドに対する完了を通知するように通信部１５０に指示する（ステップＳ１２６）。通信部１５０は、指示に応じて、ストレージ１１に受信完了のメッセージを送信する。

【００７２】

そして、ストレージ１２から準備完了のメッセージ（ステップＳ１２２において送信したwrite コマンドに対する応答）が送信されるのを待つ（ステップＳ１２７）。ストレージ１２からの準備完了のメッセージを受信したことが通信部１５０から通知されると、通信部１５０に、バッファメモリ１５２に格納されたデータをストレージ１２に向けて送信させる（ステップＳ１２８）。その後、ストレージ１２から受信完了のメッセージが送信されるのを待ち（ステップＳ１２９）、ストレージ１２からの受信完了のメッセージを受信したことが通信部１５０

から通知されると、処理を終了する。

【００７３】

通信部１５０は、外部からコマンドを受信すると、受信したコマンドを中継処理部１５１に渡す。また、外部から各種のメッセージを受信すると、受信したことを中継処理部１５１に通知する。さらに、データを受信した場合には、データを格納すべきバッファメモリ１５２の領域を中継処理部１５１に問い合わせ、バッファメモリ１５２中の指定された領域にデータを格納する。また、中継処理部１５１の指示に応じて、指定されたメッセージを指定されたストレージに送信する。さらに、バッファメモリ１５２中の領域と送信先のストレージとが指定された場合には、その領域中のデータを指定されたストレージに送信する。

【００７４】

中継処理部１５１がデータの送信先のストレージを決定する際に、あらかじめ送信先のストレージ（この例ではストレージ１２）が固定的に決められている場合には特に選択処理を行わないが、データの転送元のストレージ（この例ではストレージ１１）に応じてデータの送信先のストレージが決まる場合には、転送元のストレージに応じて送信先のストレージを選択する処理を行う。また、送信先のストレージのデータを送信するのに先立って、データの転送元のストレージから送信先が指定されることもある。なお、バッファメモリ１５２に格納されたデータは、送信先のストレージ（この例ではストレージ１２）へのデータの送信が完了すると破棄される。

【００７５】

ストレージ１２の通信部１０２は、中継装置１５がステップＳ１２２において送信したwrite コマンドを受信すると、write コマンドを処理シーケンサ１０３に渡す。処理シーケンサ１０３は、中継装置１５から送られてくるデータを格納するのに必要な領域をバッファメモリ１０６に確保する。そして、通信部１０２に、準備完了のメッセージを中継装置１５に対して送信させる。この準備完了メッセージは、中継装置１５がステップＳ１２７において待つメッセージである。その後、中継装置１５からデータが送られてくると、ストレージ１２の通信部１０２は、データを格納すべき領域を処理シーケンサ１０３に問い合わせ、バッフ

ァメモリ１０６に格納する。データを全て格納したならば、受信完了のメッセージを中継装置１５に送信する。この受信完了メッセージは、中継装置１５がステップＳ１２９において待つメッセージである。また、ストレージ１２の処理シーケンサ１０３は、バッファメモリ１０６へのデータ格納後、ＩＯスケジューラ１０４に対して、処理の種類（この場合には書き込み）と、処理の識別ＩＤと、処理の対象となる記憶媒体１０１中の領域を示す情報と、処理の対象となるバッファメモリ１０６の領域を示す情報とを登録する。ＩＯスケジューラ１０４は、書き込み要求に応じて、媒体制御部１０５に、書き込み対象のデータの書き込み指示を行う。媒体制御部１０５は、登録内容に応じてバッファメモリ１０６から記憶媒体１０１へのデータの書き込み処理を行う。

【００７６】

この実施の形態では、中継装置１５は、ストレージ１１の設置場所において地震等の災害が発生したときに、災害の影響が波及すると想定される範囲の外に設置される。また、中継装置１５は、ストレージ１１と中継装置１５との間のデータ転送時間が、ストレージ１１とストレージ１２との間で直接データを転送した場合のデータ転送時間よりも短くなるような位置に設置されている。さらに、バックアップのためにストレージ１１からストレージ１２に向けてデータを転送する際に、ストレージ１１を使用しているホスト１０は、中継装置１５へのデータ転送が完了したら、ストレージ１１からストレージ１２へのデータ転送が完了したと認識する。

【００７７】

従って、ストレージ１１が被災しても中継装置１５は被災せず、かつ、ストレージ１１から中継装置１５へのデータ転送時間が短いので、耐障害性が向上する。また、ホスト１０は、ストレージ１２にデータが格納されるのを待たずに、次の処理を開始することができるので、データ転送に伴う処理の遅れが改善される。

【００７８】

実施の形態２．

第１の実施の形態では、中継装置１５は、ストレージ１１からの１ブロックの

データを全て受信してから、ストレージ１２に対するデータの送信を開始したが、ストレージ１１からの１ブロックのデータの受信の完了を待たずに、ストレージ１２に対するデータの送信を開始してもよい。図６は、ストレージ１１からのデータの受信の完了を待たずにストレージ１２に対するデータの送信を開始する制御を行う第２の実施の形態の中継処理部１５１の動作を示すフローチャートである。なお、データ複製システムの構成およびストレージ１１，１２と中継装置１５の構成は第１の実施の形態の場合と同じである（図１～図３参照）。

【００７９】

中継処理部１５１は、第一の実施の形態に示す場合と同様に、ストレージ１１がステップＳ１０１において送信したwrite コマンドを受信する。すると、中継処理部１５１は、図６に示すように、ストレージ１１から送られてくるデータを格納するのに必要な領域をバッファメモリ１５２に確保する（ステップＳ１４０）。中継処理部１５１は、ストレージ１１からのwrite コマンドに含まれるデータ量の情報に基づいてデータの格納領域を確保する。そして、受信準備完了の通知をストレージ１１に送るように通信部１５０に指示する（ステップＳ１４１）。また、write コマンドをストレージ１２に送るように通信部１５０に指示する（ステップＳ１４２）。このwrite コマンドは、ストレージ１１から受信するデータをストレージ１２に書き込ませるためのwrite コマンドである。

【００８０】

次いで、ストレージ１１からデータが届くのを待ち（ステップＳ１４３）、データが届いて通信部１５０からデータを格納すべきバッファメモリ１５２の領域の問い合わせを受けると、ステップＳ１４０で確保した領域を通信部１５０に知らせる（ステップＳ１４４）。次いで、ストレージ１２から準備完了のメッセージが送信されるのを待ち（ステップＳ１４５）、ストレージ１２からの準備完了のメッセージを受信したことが通信部１５０から通知されると、通信部１５０に、バッファメモリ１５２に格納されたデータをストレージ１２に向けて送信させる（ステップＳ１４６）。

【００８１】

ステップＳ１４６の処理を行っているときに、ストレージ１１からのデータが

バッファメモリ１５２に格納され、バッファメモリ１５２から読み出されたデータがストレージに向けて送信されるが、通信部１５０は、データの読み出し位置（読み出しアドレス）が格納位置（格納アドレス）を追い越さないように制御する。すなわち、読み出し位置が格納位置に追いついたら、バッファメモリ１５２からのデータの読み出しを中止する。

【００８２】

そして、データのバッファメモリ１５２への格納が完了するのを待ち（ステップＳ１４７）、全てのデータがバッファメモリ１５２に格納されたことが通信部１５０から通知されると、ストレージ１１にwrite コマンドに対する完了を通知するように通信部１５０に指示する（ステップＳ１４８）。その後、ストレージ１２から受信完了のメッセージが送信されるのを待ち（ステップＳ１４９）、ストレージ１２からの受信完了のメッセージを受信したことが通信部１５０から通知されると、処理を終了する。

【００８３】

この実施の形態では、第１の実施の形態に比べて、中継装置１５からストレージ１２へのデータの送信を早めに完了させることができる。なお、ストレージ１１，１２の動作は、第１の実施の形態のそれらの動作と同じである。

【００８４】

実施の形態３．

第１および第２の実施の形態では、ストレージ１１のデータのバックアップが実現されたが、ミラーリングによってストレージ１１のデータをストレージ１２に転送するようにしてもよい。図７は、第３の実施の形態、すなわちミラーリングを行う場合のストレージ１１のストレージコントローラ１００における処理シーケンサ１０３の動作を示すフローチャートである。なお、データ複製システムの構成およびストレージ１１，１２と中継装置１５の構成は第１の実施の形態の場合と同じである（図１～図３参照）。また、ストレージ１１において、通信部１０２、ＩＯスケジューラ１０４および媒体制御部１０５の動作は、第１の実施の形態のそれらの動作と同じである。

【００８５】

第１および第２の実施の形態に示したバックアップは、ホスト１０がストレージ１１に対して複写コマンドを出力したことを契機に開始される。第３の実施の形態として示すミラーリングは、ホスト１０がストレージ１１に対してデータの書き込みを指示するwrite コマンドを出力したことを契機に開始される。

【００８６】

ストレージ１１の通信部１０２がホスト１０からwrite コマンドを受信すると、通信部１０２はそのwrite コマンドを処理シーケンサ１０３に渡す。すると、処理シーケンサ１０３は、図７に示すように、ホスト１０から受け取るデータを格納するのに必要な領域をバッファメモリ１０６に確保する（ステップＳ１６０）。また、準備完了の通知をホスト１０に送るように通信部１０２に指示する（ステップＳ１６１）。通信部１０２は、指示に応じて、準備完了の通知をホスト１０に送る。

【００８７】

そして、処理シーケンサ１０３は、ホスト１０からデータが到着するのを待ち（ステップＳ１６２）、データが届いて通信部１０２からデータを格納すべきバッファメモリ１０６の領域の問い合わせを受けると、ステップＳ１６０で確保した領域を通信部１０２に知らせる（ステップＳ１６３）。次いで、write コマンドを中継装置１５に送るように通信部１０２に指示する（ステップＳ１６４）。通信部１０２は、指示に応じて、中継装置１５にwrite コマンドを送信する。

【００８８】

次いで、処理シーケンサ１０３は、ホスト１０からのデータのバッファメモリ１０６への格納の完了を待ち（ステップＳ１６５）、全てのデータがバッファメモリ１０６に格納されたことが通信部１０２から通知されると、処理シーケンサ１０３は、ＩＯスケジューラ１０４に対して、処理の種類（この場合には書き込み）と、処理の識別ＩＤと、処理の対象となる記憶媒体１０１中の領域を示す情報と、処理の対象となるバッファメモリ１０６の領域を示す情報とを登録する。ＩＯスケジューラ１０４は、書き込み要求に応じて、媒体制御部１０５に、書き込み対象のデータの書き込み指示を行う。媒体制御部１０５は、登録内容に応じてバッファメモリ１０６から記憶媒体１０１へのデータの書き込み処理を行う（

ステップＳ１６６）。

【００８９】

そして、処理シーケンサ１０３は、中継装置１５から受信準備完了のメッセージ（ステップＳ１６４において送ったwrite コマンドに対する応答）が送信されるのを待つ（ステップＳ１６７）。中継装置１５からの受信準備完了のメッセージを受信したことが通信部１０２から通知されると、処理シーケンサ１０３は、通信部１０２に、バッファメモリ１０６に格納されたデータを中継装置１５に向けて送信させる（ステップＳ１６８）。その後、中継装置１５から受信完了のメッセージが送信されるのと、媒体制御部１０５からの書き込み完了通知とを待ち（ステップＳ１６９）、中継装置１５からの受信完了のメッセージを受信したことが通信部１０２から通知され、かつ、媒体制御部１０５からの書き込み完了通知を受けると、ホスト１０に書き込みの完了を通知し（ステップＳ１７０）、処理を終了する。ホスト１０は、ストレージ１１から完了通知を受けたら、すなわち、中継装置１５へのデータ転送および記憶媒体１０１へのデータ書き込みが完了したら、ミラーリングが完了したと認識する。

【００９０】

中継装置１５がストレージ１１から送られるデータを中継する動作や、ストレージ１２が中継装置１５から送られるデータを記憶媒体１０１に記憶させる動作は、第一の実施の形態と同様である。

【００９１】

この実施の形態でも、ストレージ１１が被災しても中継装置１５は被災せず、かつ、ストレージ１１から中継装置１５へのデータ転送時間が短いので、耐障害性が向上する。また、ホスト１０がwrite コマンドを送信してから次の処理を開始するまでの時間を短縮化できる。

【００９２】

実施の形態４．

図８は、本発明によるデータ複製システムの第４の実施の形態を示すブロック図である。図８に示すデータ複製システムにおいて、ストレージ１１が、ストレージ１１を使用するホスト１０とローカルに接続されている。ストレージ１１は

、ネットワーク１３を介して中継装置１５－１～１５－ｎに接続されている。また、中継装置１５－１～１５－ｎは、ネットワーク１４を介してストレージ１２に接続されている。なお、ストレージ１１，１２の構成は、第１の実施の形態のストレージ１１，１２の構成と同じであり、中継装置１５－１～１５－ｎの構成は、第１の実施の形態の中継装置１５の構成と同じである。また、各中継装置１５－１～１５－ｎがストレージ１１から受信したデータをストレージ１２に転送する動作や、ストレージ１２が受信したデータを記憶媒体１０１に記憶させる動作は、第１の実施の形態の中継装置１５やストレージ１２の動作と同様である。

【００９３】

第一の実施の形態と同様に、ストレージ１１は主体的にデータを送信する。すなわち、ホスト１０がストレージ１１からデータを読み出して中継装置１５－１～１５－ｎにデータを送信するのではなく、ストレージ１１が直接中継装置１５にデータを送信する。ストレージ１２も、中継装置１５から直接データを受信する。また、中継装置１５－１～１５－ｎは、ストレージ１１，１２の設置場所において地震等の災害が発生したときに、災害の影響が波及すると想定される範囲の外に設置される。さらに、ストレージ１１またはストレージ１２と中継装置１５－１～１５－ｎとの間のデータ転送時間が、ストレージ１１とストレージ１２との間で直接データを転送した場合のデータ転送時間よりも短くなるように、中継装置１５－１～１５－ｎが設置される。

【００９４】

次に、データ複製システムの動作を説明する。図９は、ストレージ１１のストレージコントローラ１００における処理シーケンサ１０３の動作を示すフローチャートである。ここでは、データの転送元としてのストレージ１１が、データの転送先としてのストレージ１２に向けてデータを転送して、バックアップを行う場合の例を説明する。

【００９５】

処理シーケンサ１０３は、通信部１０２を介してホスト１０から複写コマンドを受け取ると、図９に示すように、まず、複写対象のデータの範囲の先頭のブロックをデータ転送対象のブロックとして設定する（ステップＳ２００）。次いで

、中継装置１５－１～１５－ｎのうちから使用する中継装置を選択する（ステップＳ２０１）。ここでは、中継装置１５－１が選択されたとする。そして、通信部１０２に、選択した中継装置１５－１に対してwrite コマンドを送信するように指示する（ステップＳ２０２）。また、データ転送対象のデータの読み出し要求をＩＯスケジューラ１０４に登録する（ステップＳ２０３）。ステップＳ２０３において読み出し要求が登録されたときのＩＯスケジューラ１０４と媒体制御部１０５の動作は、第１の実施の形態で説明したステップＳ１０２における動作と同様である。

【００９６】

そして、処理シーケンサ１０３は、通信部１０２を介して入力される中継装置１５－１からの受信準備完了のメッセージと、媒体制御部１０５からの読み出し完了通知との双方を待ち（ステップＳ２０４）。双方が入力されたら、バッファメモリ１０６に記憶されたデータを中継装置１５－１に転送するように通信部１０２に指示する（ステップＳ２０５）。そして、処理シーケンサ１０３は、通信部１０２を介して入力される中継装置１５－１からの受信完了のメッセージを待つ（ステップＳ２０６）。

【００９７】

通信部１０２は、ステップＳ２０３においてバッファメモリ１０６に記憶させたデータを中継装置１５－１に送信し、中継装置１５－１から受信完了のメッセージを受けたら、受信完了のメッセージを処理シーケンサ１０３に出力する。処理シーケンサ１０３は、受信完了のメッセージを入力すると、複写対象の全てのデータの中継装置１５への転送が完了したか否か確認する（ステップＳ２０７）。完了していなければ、複写対象のデータの範囲の次のブロックをデータ転送対象のブロックとして設定し（ステップＳ２０９）、ステップＳ２０１に戻る。

【００９８】

複写対象の全てのデータの転送が完了していれば、処理シーケンサ１０３は、ホスト１０に対して完了通知を出力するように通信部１０２に指示し（ステップＳ２０８）、処理を終了する。

【００９９】

なお、１ブロックのデータ量は、あらかじめシステムに設定されている量である。また、処理シーケンサ１０３は、１ブロックのデータ量を、転送先のストレージ１２のバッファメモリ１０６および中継装置１５のバッファメモリ１５２の容量に応じて変化させるようにしてもよい。

【０１００】

ステップＳ２０１での中継装置を選択する方法として、例えば、中継装置１５－１～１５－ｎを順に選択したり、負荷の軽い（他のデータ転送に用いられていない）中継装置を選択したり、乱数を用いて選択したりする方法がある。

【０１０１】

あるいは、複数の中継装置を同時に使用してデータ転送を並列に実行してもよい。図１０は、複数の中継装置を同時に使用する場合のストレージ１１の動作を示すフローチャートである。ここでは、１ブロックのデータを転送する場合を例にする。

【０１０２】

ストレージ１１において、処理シーケンサ１０３が、複写対象のデータの範囲の先頭のブロックをデータ転送対象のブロックとして設定すると（ステップＳ２２０）、ストレージ１１は、全中継装置１５－１～１５－ｎに対してデータを転送する（ステップＳ２２１）。

【０１０３】

図１１は、ステップＳ２２１の処理を具体的に示すフローチャートである。処理シーケンサ１０３は、個々の中継装置１５－１～１５－ｎに対して、それぞれ、図１１に示すフローチャートに従ってデータを送信する。ここでは、中継装置１５－１にデータを送信する場合を例に説明する。処理シーケンサ１０３は、バッファメモリ１０６に格納されているデータから、中継装置１５－１に送信するデータを選択する（ステップＳ２４０）。そして、その中継装置１５－１に対してwrite コマンドを送るように通信部１０２に指示する（ステップＳ２４１）。また、ＩＯスケジューラ１０４に、転送対象のデータの読み出し要求を登録する（ステップＳ２４２）。ステップＳ２４２において読み出し要求が登録されたときのＩＯスケジューラ１０４と媒体制御部１０５の動作は、第１の実施の形態で

説明したステップＳ１０２における動作と同様である。処理シーケンサ１０３は、通信部１０２を介して入力される中継装置１５－１からの受信準備完了のメッセージと、媒体制御部１０５からの読み出し完了通知との双方を待ち（ステップＳ２４３）。双方が入力されたら、バッファメモリ１０６に記憶されたデータを中継装置１５－１に送信するように通信部１０２に指示する（ステップＳ２４４）。ここでは中継装置１５－１にデータを送信する場合を例に説明したが、処理シーケンサ１０３は、ステップＳ２４０～Ｓ２４４の処理を全ての中継装置１５－１～１５－ｎについて実行する。

【０１０４】

さらに、処理シーケンサ１０３は、いずれかの中継装置から受信完了のメッセージを受けたことを通信部１０２から通知されたら（ステップＳ２２２）、未転送のデータがまだあるか否か確認し（ステップＳ２２３）、ある場合には、受信完了のメッセージを送信した中継装置に対して、write コマンドを送るように通信部１０２に指示する（ステップＳ２２４）。なお、ステップＳ２２４の具体的な処理は、図１１に示された処理である。

【０１０５】

ステップＳ２２３において未転送のデータがないことを確認したら、処理シーケンサ１０３は、各中継装置から受信完了のメッセージを受けるのを待ち（ステップＳ２２５）、データを転送した全ての中継装置から受信完了のメッセージを受けたら（ステップＳ２２６）、ホスト１０に対して完了通知を出力するように通信部１０２に指示し（ステップＳ２２７）、処理を終了する。

【０１０６】

図１２は、１ブロックのデータを５つに分け、３台の中継装置１５－１，１５－２，１５－３を使用する場合のデータ転送の例を示すタイミング図である。ストレージ１１は、ホスト１０からの複写コマンドに応じて、中継装置１５－１，１５－２，１５－３のそれぞれにwrite コマンドを送信する。そして、受信準備完了のメッセージを送信した中継装置１５－１，１５－２，１５－３にデータ①，②，③を転送する。

【０１０７】

図１２に示す例では、中継装置１５－２，１５－３が先に完了通知を送信したので、ストレージ１１は、中継装置１５－２，１５－３に対してwrite コマンドを送信し、その後、データ④，⑤を送信する。そして、全ての中継装置１５－１，１５－２，１５－３からの完了通知の受信を確認した時点で、ストレージ１１は、ホスト１０に完了通知を出力する。

【０１０８】

この実施の形態では、複数の経路を用いてデータ転送を実行することができる。よって、個々の経路の使用率および中継装置の処理能力の影響を受けにくくすることができる。例えば、図１２に示された例では、中継装置１５－１への経路が輻輳していたか、または中継装置１５－１の処理能力が低かったことになるが、他の中継装置１５－２，１５－３を用いることによって、中継装置１５－１への経路の輻輳または中継装置１５－１の処理能力が低いことは、ストレージ１１からのデータ転送に対して大きな影響を与えることはない。

【０１０９】

ここでは、各中継装置１５－１～１５－ｎに、それぞれ異なるデータを送信する場合について説明したが、同一のデータのwrite コマンドを各中継装置１５－１～１５－ｎに送信するようにしてもよい。例えば、送信すべきデータとしてデータＡ～Ｃがあるとする。ストレージ１１は、データＡ～Ｃのいずれのwrite コマンドについても、各中継装置１５－１～１５－ｎに送信する。

【０１１０】

この場合、ストレージ１１は、各中継装置１５－１～１５－ｎに送信する同一のデータの組毎に識別子を設け、その識別子を各write コマンドに付加する。識別子としては、送信される各データの組の順番を表す数値を用いればよい。例えば、最初にデータＡを各中継装置に送信するのであれば、各中継装置に送信する各write コマンドに一番目を表す「１」という識別子を付加する。続いて、各中継装置にデータＢを送信するのであれば、各中継装置に送信する各write コマンドに二番目を表す「２」という識別子を付加する。他のデータのwrite コマンドにも同様に識別子を付加する。

【０１１１】

各中継装置１５－１～１５－ｎは、ストレージ１１に対してデータの転送完了を通知するときには、write コマンドに付加された識別子も通知する。従って、ストレージ１１は、同一の識別子が付加された応答を各中継装置１５－１～１５－ｎから受信する。ストレージ１１は、応答に付加された識別子が初めて受信する識別子であれば、ホスト１０にそのデータの転送終了を通知する。その後、同一の識別子が付加された応答を受信した場合、その応答に対しては何も処理を行わない。

【０１１２】

また、各中継装置１５－１～１５－ｎは、識別子を付加したままwrite コマンドを待機系のストレージ１２に送信する。従って、ストレージ１２は同一の識別子が付加されたwrite コマンドを各中継装置１５－１～１５－ｎから受信する。ストレージ１２は、write コマンドに付加された識別子が初めて受信する識別子であれば、そのwrite コマンドに従って処理を進める。その後、同一の識別子が付加されたwrite コマンドを受信した場合には、そのwrite コマンドを送信した中継装置に対して、転送が完了した通知のみを行う。

【０１１３】

このように同一データのwrite コマンドを各中継装置に送信すれば、複数のデータ転送経路のうち最も速度が速い経路を用いて処理を進めることができる。従って、ストレージ１１がミラーリングやバックアップを行う際、中継装置からの応答の待ち時間を短くすることができる。その結果、ホスト１０が次の処理を開始するまでの時間を短縮化される。

【０１１４】

第一の実施の形態から第四の実施の形態において、中継装置１５に不揮発性記憶装置を設け、中継装置１５が、ストレージ１１から受信したコマンドおよびデータをその不揮発性記憶装置に記録し、不揮発性記憶装置に記録したコマンドを任意のタイミングでストレージ１２に発行するようにしてもよい。中継装置１５に設ける不揮発性記憶装置としては、例えば、単体の磁気ディスク装置、光ディスク装置または光磁気ディスク装置の集合であるディスクアレイ装置等を使用すればよい。あるいは、バッテリバックアップされたメモリを使用してもよい。

【０１１５】

中継装置１５は、例えば、中継装置１５およびストレージ１２が接続されるネットワーク１４の通信量が所定のしきい値以下になった場合等に、ストレージ１２に対してwrite コマンドを発行すればよい。write コマンド発行のタイミングは、不揮発性記憶装置にコマンドおよびデータを記録してから所定の時間が経過したとき、所定の時刻になったとき、あるいはストレージ１２からwrite コマンドの発行を要求されたとき等であってもよい。

【０１１６】

このように不揮発性記憶装置にストレージ１１からのコマンドおよびデータを記録すれば、コマンドやデータが喪失されることがない。そして、ストレージ１１からのwrite コマンド受信と連動せずに、任意のタイミングでストレージ１２にwrite コマンドを発行できる。その結果、ストレージ１２を常時稼働させておかなくてもよくなる。同様に、中継装置１５およびストレージ１２が接続されるネットワーク１４も常時接続状態を保つ必要がなくなり、ネットワーク１４の運用コストも低減される。

【０１１７】

また、通常は、第一の実施の形態から第四の実施の形態に示したようにストレージ１１からのwrite コマンド受信と連動させてストレージ１２にwrite コマンドを発行し、ネットワーク１４の通信量が所定のしきい値以上になったときに、コマンドやデータを不揮発性記憶装置に記録するようにしてもよい。この場合、ネットワーク１４の通信量を平均化することが可能となる。また、ネットワーク１４として運用コストが安価な回線を使用することができ、データ複製システム全体のコストを低減することができる。

【０１１８】

第一の実施の形態から第四の実施の形態において、データ転送処理手段は、処理シーケンサ１０３および通信部１０２によって実現される。中継処理手段は、中継処理部１５１および通信部１５０によって実現される。

【０１１９】

実施の形態５.

図１３は、本発明によるデータ複製システムの第５の実施の形態を示すブロック図である。図１３に示すデータ複製システムにおいて、ストレージ２０が、ストレージ２０を使用するホスト１０とローカルに接続されている。ストレージ２０は、ネットワーク１３を介してストレージ２１に接続されている。

【０１２０】

ストレージ２０，２１は、例えば、単体の磁気ディスク装置、光ディスク装置または光磁気ディスク装置である。ストレージ２０，２１として、単体の磁気ディスク装置、光ディスク装置または光磁気ディスク装置の集合であるディスクアレイ装置を使用することもできる。ホスト１０とストレージ２０とは、ＳＣＳＩ、ファイバチャネル（Fibre channel ）、イーサネット（登録商標）等で接続される。

【０１２１】

図１４は、図１３に示すストレージ２０の構成例を示すブロック図である。なお、ストレージ２１の構成も、図１４に示すような構成である。図１４に示すように、ストレージ２０は、ストレージコントローラ２００とストレージ本体である記憶媒体１０１とを含む。ストレージコントローラ２００は、ホスト１０および他のストレージと通信を行う通信部２０４、各処理のシーケンスを管理する処理シーケンサ２０１、記憶媒体１０１に対する処理命令の順序制御を行うＩＯスケジューラ１０４、ＩＯスケジューラ１０４が発行する処理命令に従って記憶媒体１０１の動作を制御する媒体処理部１０５、ホスト１０等から記憶媒体１０１へのデータおよび記憶媒体１０１からホスト１０等へのデータを一時記憶するバッファメモリ１０６、他のストレージに送るデータを冗長化する冗長化部２０２、および他のストレージから送られてきた冗長化データから元のデータを復元する復元部２０２を含む。処理シーケンサ１０３は、例えば、プログラムに従って動作するＣＰＵで実現される。

【０１２２】

次に、実施の形態５のデータ複製システムの動作を、図１５および図１６のフローチャートを参照して説明する。図１５は、ストレージコントローラ２００における処理シーケンサ２０１の動作を示すフローチャートであり、図１６は通信

部２０４の動作を示すフローチャートである。

【０１２３】

処理シーケンサ２０１は、通信部２０４を介してホスト１０から複写コマンドを受け取ると、図１５に示すように、まず、複写対象のデータの範囲の先頭のブロックをデータ転送対象のブロックとして設定する（ステップＳ３００）。次いで、通信部２０４に、複写コマンドで指定された複写先のストレージに対して冗長化write コマンドを送信するように指示する（ステップＳ３０１）。冗長化write コマンドは、冗長化されたデータに基づく書き込みを指示するコマンドである。冗長化write コマンドには、ブロックサイズすなわちデータ量を示す情報とともに、データが冗長化されることを示す情報が含まれている。また、処理シーケンサ２０１は、データ転送対象のデータの読み出し要求をＩＯスケジューラ１０４に登録する。ＩＯスケジューラ１０４は、読み出し要求に応じて、媒体制御部１０５に、データ転送対象のデータの読み出し指示を行う。媒体制御部１０５は、読み出し指示に従って、データ転送対象のデータを記憶媒体１０１からバッファメモリ１０６に出力させる（ステップＳ３０２）。

【０１２４】

そして、処理シーケンサ２０１は、通信部２０４を介して入力されるストレージ２１からの受信準備完了のメッセージと、媒体制御部１０５からの読み出し完了通知との双方を待ち（ステップＳ３０３）。双方が入力されたら、バッファメモリ１０６に格納されたデータの冗長化を冗長化部２０２に行わせる（ステップＳ３０４）。冗長化は、元のデータ（以下、元データという。）から冗長データを作成することによって行う。以下の説明において、元データおよび新たに作成された冗長データの集合を冗長化されたデータ群と記す。処理シーケンサ２０１は、冗長化されたデータ群をストレージ２１に転送するように通信部２０４に指示する。通信部２０４は、この指示に応じて、冗長化されたデータ群をストレージ２１に送信する（ステップＳ３０５）。

【０１２５】

ただし、通信部２０４は、元データと冗長データとを別々のデータのまとまりとしてストレージ２１に送信する。この「データのまとまり」とは、各種データ

転送プロトコルの最小のデータ送信単位のことである。「データのまとまり」の具体例としては、例えば、ＴＣＰやＵＤＰにおけるセグメント、インターネットプロトコルにおけるパケット、あるいはファイバチャネルプロトコルやイーサネット（登録商標）の物理層におけるフレーム等がある。従って、例えばインターネットプロトコルに従って送信を行う場合、通信部２０４は、元データと冗長データとを別々のパケットで送信する。なお、一般に、パケット等のデータ送信単位の内部には、伝送中に生じたエラー（データ化け）を検出するためのエラー検出データが付加される。ステップＳ３０４で作成される冗長データは、独立したデータのまとまりとして送信されるデータであり、パケット等に付加されるエラー検出データとは異なるデータである。

【０１２６】

また、後述するように、元データを分割し、分割された元データに基づいて冗長データを作成する。ステップＳ３０５において、通信部３０５は、分割された元データをそれぞれ個々のデータ送信単位として送信する。例えば、元データをＮ個に分割し、Ｎ個のデータからｍ個の冗長データを作成したとする。この場合、通信部は元データおよび冗長データをそれぞれＮ個、ｍ個のパケット（あるいはセグメント、フレーム等）で送信する。通信部２０４が元データと冗長データとをそれぞれ別々のデータのまとまりとして送信したならば、通信部２０４を介して入力されるストレージ２１からの受信完了のメッセージを待つ（ステップＳ３０６）。

【０１２７】

通信部２０４は、冗長化されたデータ群（元データと冗長データ）を全てストレージ２１に送信し、ストレージ２１から受信完了のメッセージを受けたら、受信完了のメッセージを処理シーケンサ２０１を出力する。処理シーケンサ２０１は、受信完了のメッセージを入力すると、複写対象の全てのデータの中継装置１５への転送が完了したか否か確認する（ステップＳ３０７）。完了していなければ、複写対象のデータの範囲の次のブロックをデータ転送対象のブロックとして設定し（ステップＳ３０９）、ステップＳ３０１に戻る。

【０１２８】

複写対象の全てのデータの転送が完了していれば、処理シーケンサ２０１は、ホスト１０に対して完了通知を出力するように通信部２０４に指示し（ステップＳ３０８）、処理を終了する。

【０１２９】

なお、１ブロックのデータ量は、あらかじめシステムに設定されている量である。また、処理シーケンサ２０１は、１ブロックのデータ量を、転送先のストレージ２１のバッファメモリ１０６の容量に応じて変化させるようにしてもよい。

【０１３０】

通信部２０４は、外部から入力されたデータが、コマンド、受信準備完了のメッセージ、受信完了のメッセージ等の制御系メッセージであった場合には、入力されたデータを処理シーケンサ２０１に渡す。また、ストレージ２０に書き込まれるべきデータであった場合には、データを格納すべき場所を処理シーケンサ２０１に問い合わせる。そして、処理シーケンサ２０１から指定されたバッファメモリ１０６中の領域にデータを格納するか、または、復元部２０３にデータを復元させた後、指定されたバッファメモリ１０６中の領域にデータを格納する。

【０１３１】

また、通信部２０４は、処理シーケンサ２０１から指定されたコマンドまたは完了通知を、指定されたストレージまたはホストに送信する処理も行う。さらに、処理シーケンサ２０１から指定されたバッファメモリ１０６中のデータを、指定されたストレージまたはホストに送信する処理も行う。また、冗長化部２０２から冗長化されたデータを受け取り、受け取ったデータを、指定されたストレージまたはホスト１０に送信する処理も行う。

【０１３２】

なお、通信部２０４は、他のストレージにコマンドを送信する場合に、各コマンドを区別するためのコマンド識別子をコマンドに付加する。コマンド識別子の例として、コマンドを送信する毎に１加算した数値である発行番号（後述する発行ＩＤ）がある。また、冗長化されたデータ群を送信する際に、その送信に関連したコマンドのコマンド識別子を付加する。例えば、冗長化write コマンドを送信するときに、その冗長化write コマンドを識別するためのコマンド識別子を付

加する。そして、その冗長化write コマンドに対応する冗長化されたデータ群（元データおよび冗長データ）にも、冗長化write コマンドと同じコマンド識別子を付加する。

【０１３３】

さらに、通信部２０４は、各コマンドやデータに、送信元となるストレージの識別子も付加する。

【０１３４】

次に、冗長化部２０２の動作について説明する。冗長化部２０２は、指定されたデータをバッファ１０６から取得し、規定された冗長化方法を用いて、冗長化されたデータ群を作成する。冗長化部２０２は、指定された元データをＮ個のデータに分割する。そして、Ｎ個に分割された元データからｍ個の冗長データを作成する。Ｎ，ｍは自然数である。このＮ＋ｍ個のデータ群が、冗長化されたデータ群となる。冗長化されたデータ群には、復元時に使用するための識別情報が付加される。

【０１３５】

以下、冗長化の具体例について説明する。冗長化部２０２は、例えば、元データをその先頭からＮ等分するなどの方法により、元データをＮ個に分割する。この方法を採用した場合、先頭から何番目のデータであるのかを示す番号を識別情報として使用する。冗長化部２０２は、元データをＮ個のデータに分割したならば、ＲＡＩＤ３およびＲＡＩＤ５の装置等と同様にパリティ演算によってパリティデータを作成し、そのパリティデータを冗長データとする。すなわち、Ｎ個に分割された各データおよび冗長データにおいて、対応する各ビットの値の総和が必ず奇数（または必ず偶数）になるように冗長データを作成する。この場合、１個の冗長データを作成すればよい。このように冗長ビットを作成すれば、分割後のＮ個の元データのうち、一つが欠落しても、欠落したデータを復元できる。

【０１３６】

冗長化の方法は上記のパリティ演算のみに限定されず、冗長データの数も１個とは限らない。例えば、ダブルパリティ演算によって冗長データを作成してもよい。また、元データに基づくＥＣＣ（Error Correcting Code：誤り訂正符号）

を冗長データとして用いてもよい。このように様々な冗長化の方法があるが、以下の説明では、パリティ演算によって１個の冗長データを作成した場合を例に説明する。

【０１３７】

次に、データの転送先であるストレージ２１の動作を説明する。ストレージ２０からストレージ２１のストレージコントローラ２００における通信部２０４に冗長化write コマンドが送られる。通信部２０４は、冗長化write コマンドを受信すると、処理シーケンサ２０１に冗長化write コマンドを渡し、冗長化write コマンドにもとづく処理（冗長化write 処理）の開始を指示する。

【０１３８】

図１６は、ストレージ２１における処理シーケンサ２０１の動作を示すフローチャートである。処理シーケンサ２０１は、複数の冗長化write 処理を並行して行うことが可能であり、コマンドを発行処理したストレージの識別子と、コマンドに付加されたコマンド識別子を用いて各処理を識別する。

【０１３９】

冗長化write 処理では、処理シーケンサ２０１は、まず、送られてくるデータを記憶できる領域をバッファ１０６中に確保する（ステップＳ３２０）。次いで、通信部２０４に冗長化write コマンドを送信したストレージ２０に向けて受信準備の完了通知を送信させる（ステップＳ３２１）。そして、ストレージ２０からデータが届くのを待ち（ステップＳ３２２）、データが到着したことが通信部２０４から通知されたら、データ転送元のストレージ２０の識別子とデータに付加されているコマンド識別子とを指定し、データを復元部２０３に送るように通信部２０４に指示する（ステップＳ３２３）。

【０１４０】

次いで、処理シーケンサ２０１は、冗長化write コマンドに対応して送られてきたデータの個数を判断する（ステップＳ３２４）。各データには、冗長化write コマンドに対応したコマンド識別子が付加されているので、処理シーケンサ２０１は、受信した各データが冗長化write コマンドに対応して送られてきたデータか否かを確認することができる。データの個数が所定の個数（ｎ個）未満の場

合にはステップＳ３２２に移行し、所定の個数（ｎ個）である場合にはステップＳ３２５に移行する（ステップＳ３２４）。この所定の個数は、元データを復元することができるデータの個数であり、冗長化の方式によって異なる。例えば、Ｎ個に分割された元データからパリティ演算によって１個の冗長データが作成された場合、ストレージ２０からはＮ＋１個のデータが送られてくる。この場合、Ｎ個のデータを受信すれば、元データを復元することができる。従って、ステップＳ３２４では受信したデータの数がＮ個未満か否かを判断すればよい。

【０１４１】

ステップＳ３２５において、処理シーケンサ２０１は、データ転送元のストレージ２０の識別子およびコマンド識別子と、ステップＳ３２０で確保したバッファ１０６中の領域とを指定し、データの復元を復元部２０３に指示する。そして、データ転送元のストレージ２０に応答を送信するように通信部２０４に指示する（ステップＳ３２６）。また、ＩＯスケジューラ１０４に対して、ステップＳ３２０で確保したバッファ１０６中の領域中のデータを、冗長化write コマンドで指定された媒体中の領域に書き込むように指示を出し（ステップＳ３２７）、処理を終了する。ＩＯスケジューラ１０４にこの指示が出された場合のＩＯスケジューラ１０４および媒体制御部１０５の動作は、第１の実施の形態で説明したステップＳ１０２における動作と同様である。

【０１４２】

なお、処理シーケンサ２０１は、ステップＳ３２５の処理開始以降に通信部２０４に到着した冗長化write コマンドに関連したデータについては、通信部２０４にそのデータの破棄を指示する。例えば、Ｎ個のデータから元データを復元できる場合、Ｎ＋１個目に到着したデータは必要がない。必要なデータが到着した後にストレージ２１に到着したデータは破棄する。

【０１４３】

次に、復元部２０３の処理について説明する。復元部２０３は、バッファメモリを有し、データ転送元のストレージ２０の識別子およびコマンド識別子とともに渡されたデータをバッファメモリに蓄える。データ転送元のストレージ２０の識別子およびコマンド識別子と、バッファメモリ１０６中の復元したデータを格

納する領域とが指定され、復元が指示されると、復元部２０３は、内部のバッファから、データ転送元のストレージ２０の識別子およびコマンド識別子に該当するデータを集める。そして、各データに付加された識別情報を元に、データ群から冗長化前のデータを復元する。

【０１４４】

例えば、ＲＡＩＤ３およびＲＡＩＤ５の装置等のように、Ｎ個に分割された元データからパリティ演算によって１個の冗長データを作成して冗長化を行っていたとする。そして、分割されたＮ個の元データのうちｋ番目のデータがストレージ２１に到着せず、他のＮ－１個のデータと１個の冗長データ（合計Ｎ個のデータ）が到着したとする。この場合、ｋ番目のデータと到着したＮ個のデータにおいて、対応する各ビットの値の総和が必ず奇数（または必ず偶数）になるようにすることによって、ｋ番目のデータを作成することができる。そして、作成したｋ番目のデータと、ストレージ２１に到着した他のデータとによって、分割前の元データを復元すればよい。分割された各データには、識別情報（例えば、先頭から何番目のデータであるのかを示す番号）が付加されているので、分割された状態から分割前の元データに復元させることができる。また、Ｎ個に分割された元データが全てストレージ２１に到着し、冗長データが到着しなかった場合には、到着したＮ個のデータから分割前の元データを復元すればよい。

【０１４５】

復元部２０３は、指定されたバッファメモリ１０６中の領域に復元したデータを格納する。なお、復元部２０３の内部のバッファにおける復元処理で使用されていたデータを格納していた領域は、復元処理終了後他のデータの格納に再利用される。

【０１４６】

既述のように、冗長化の方法は、パリティ演算によって１個の冗長データを作成する方法に限られない。転送先のストレージ２１では、冗長化の方法に対応した方法で復元を行えるように制御すればよい。例えば、ＥＣＣを冗長データとしたならば、ＥＣＣに対応した方法で復元を行うように制御すればよい。

【０１４７】

また、ここでは、元データをＮ個に分割して冗長データを作成する場合を例示に説明したが、ストレージ２０の冗長化部２０２は、ステップＳ３０４において、元データを分割せずに元データと同一のデータを複製し、その複製したデータを冗長データとしてもよい。そして、ステップＳ３０５では、元データを一つのデータのまとまり（例えば、一つのパケット、セグメントあるいはフレーム等）として送信し、元データの複製である冗長データも一つのデータのまとまりとして送信してもよい。元データと冗長データとは同一のデータであり、また、分割されていないので、転送先ストレージ２１では、いずれか一方を受信したときに、受信したデータをそのまま記憶媒体１０１に書き込んでよい。また、遅れて受信したデータは破棄してよい。従って、ストレージ２１の処理シーケンサ２０１は、一回目にステップＳ３２４に移行したならば、すぐに次のステップＳ３２５に移行してよい。さらに、ステップＳ３２５において復元部２０３が復元処理を行うことなく、次のステップＳ３２６に移行してよい。

【０１４８】

このように元データと冗長データとを同一のデータとして、別々の二つのデータのまとまりとして送信すれば、一方のデータが送信過程で失われてもストレージ２１は、送信元が送信しようとした元データをそのまま受信することができる。

【０１４９】

また、ここでは、元データを分割せずに元データと同一のデータを複製し、その複製したデータを冗長データとする場合を示した。ストレージ２０の冗長化部２０２は、冗長データを作成するときに、元データをＮ個に分割し、分割したＮ個のデータについてそれぞれ同一のデータを複製してもよい。この場合、分割後のＮ個のデータの複製であるＮ個のデータが冗長データとなる。そして、ステップＳ３０５では、元データをＮ個のデータのまとまり（例えば、Ｎ個のパケット、セグメントあるいはフレーム等）として送信し、冗長データもＮ個のデータのまとまりとして送信してもよい。この場合、ストレージ２１には２×Ｎ個のデータが送信される。ストレージ２１は、この２×Ｎ個のデータのうち、元データを復元することができる一組のデータ（分割後の第一番目から第Ｎ番目までの各デ

ータ）について受信したならば、そのデータに基づいて元データを復元する。すなわち、ストレージ２１の処理シーケンサ２０１は、ステップＳ３２４において、分割後の第一番目のデータから第Ｎ番目のデータまでを全て受信したか否かを確認すればよい。そして、全て受信していればステップＳ３２５に移行し、まだ受信していないデータがあればステップＳ３２２に戻ればよい。

【０１５０】

このように複製データを冗長データとして送信すれば、２×Ｎ個のデータの一部が送信過程で廃棄されても、ストレージ２１が分割後の第一番目のデータから第Ｎ番目のデータまでを全て受信した時点でデータを復元することができる。なお、ストレージ２０の冗長化部２０２は、元データを分割してからその分割後の各データの複製を作成してもよいし、また、分割前の元データと同一のデータを複製し、元データおよび複製データをそれぞれ分割してもよい。

【０１５１】

第５の実施の形態によれば、冗長データを作成し、元データと冗長データとを別々に送信する。従って、データの一部が送信過程で失われてしまったとしても、元データを復元することができる。あるいは、送信元が送信しようとした元データをそのまま受信することができる。この結果、元データの一部が送信過程で失われた場合、転送元のストレージ２０に元データが届いていないことを通知して、ストレージ２０に再度送信させる必要がなくなり、データの転送時間を短縮することができる。

【０１５２】

図１３に示すデータ複製システムでは、冗長化されたデータ群に含まれる元データおよび冗長データは、いずれもネットワーク１３を介して送受信される。元データと冗長データとを別々のネットワークを介して送信する構成であってもよい。図１７は、元データと冗長データとを別々のネットワークを介して送信するデータ複製システムの構成例を示すブロック図である。図１７に示す構成例において、ストレージ２０は、ネットワーク１３を介してストレージ２１に接続され、またネットワーク１４を介してもストレージ２１に接続されている。ストレージ２０とストレージ２１とが二つのネットワーク１３，１４を介して接続されて

いる点以外は、図１３に示す構成と同様である。また、図１７に示すストレージ２０，２１の構成は、図１４に示す構成と同様である。ただし、通信部２０４は、二つのネットワーク１３，１４に接続される。

【０１５３】

また、図１７に示すストレージ２０の動作は、図１３に示すストレージ２０の動作と同様である。ただし、ステップＳ３０５（図１５参照）において、通信部２０４が冗長化されたデータ群をストレージ２１に送信する場合、通信部２０４は、元データと冗長データをそれぞれ別々のネットワークを介して送信する。例えば、元データを一つのデータ転送単位として（例えば一つのパケット等として）、ネットワーク１３ａを介してストレージ２１に送信し、元データの複製である冗長データを、ネットワーク１４を介してストレージ２１に送信する。

【０１５４】

このような構成によれば、ネットワーク１３，１４のいずれか一方で障害が発生したり、いずれか一方のネットワークの転送速度が遅かったとしても、もう一方のネットワークを介して元データまたは元データと同一の冗長データを送信することができる。従って、元データの複製を冗長データとする場合に、耐障害性をより高めることができる。なお、元データをＮ個に分割して冗長データを作成する場合であっても、分割した元データと冗長データとをそれぞれ別のネットワークを介して送信するようにしてもよい。

【０１５５】

実施の形態６．

第５の実施の形態では、ストレージ２０のデータのバックアップが実現されたが、ミラーリングによってストレージ２０のデータをストレージ２１に転送するようにしてもよい。図１８は、第６の実施の形態、すなわちミラーリングを行う場合のストレージ２０のストレージコントローラ２００における処理シーケンサ２０１の動作を示すフローチャートである。なお、データ複製システムの構成およびストレージ２０，２１の構成は第５の実施の形態の場合と同じである（図１３，図１４参照）。また、ストレージ２０において、通信部２０４、ＩＯスケジューラ１０４および媒体制御部１０５の動作は、第５の実施の形態のそれらの動

作と同じである。

【０１５６】

ストレージ２０は、ホスト１０からwrite コマンドを受信したときにミラーリングを開始する。ストレージ２０の通信部２０４がホスト１０からwrite コマンドを受信すると、通信部２０４はそのwrite コマンドを処理シーケンサ２０４に渡す。すると、処理シーケンサ２０１は、図１８に示すように、ホスト１０から受け取るデータを格納するのに必要な領域をバッファメモリ１０６に確保する（ステップＳ３４０）。また、準備完了の通知をホスト１０に送るように通信部２０４に指示する（ステップＳ３４１）。通信部２０４は、指示に応じて、準備完了の通知をホスト１０に送る。

【０１５７】

そして、処理シーケンサ２０１は、ホスト１０からデータが到着するのを待ち（ステップＳ３４２）、データが届いて通信部２０４からデータを格納すべきバッファメモリ１０６の領域の問い合わせを受けると、ステップＳ３４０で確保した領域を通信部２０４に知らせる（ステップＳ３４３）。次いで、冗長化writeコマンドをストレージ２１に送るように通信部２０４に指示する（ステップＳ３４４）。通信部２０４は、指示に応じて、ストレージ２１に冗長化write コマンドを送信する。

【０１５８】

次いで、処理シーケンサ２０１は、ホスト１０からのデータのバッファメモリ１０６への格納の完了を待ち（ステップＳ３４５）、全てのデータがバッファメモリ１０６に格納されたことが通信部２０４から通知されると、処理シーケンサ２０１は、ＩＯスケジューラ１０４に対して、処理の種類（この場合には書き込み）と、処理の識別ＩＤと、処理の対象となる記憶媒体１０１中の領域を示す情報と、処理の対象となるバッファメモリ１０６の領域を示す情報とを登録する。ＩＯスケジューラ１０４は、書き込み要求に応じて、媒体制御部１０５に、書き込み対象のデータの書き込み指示を行う。媒体制御部１０５は、登録内容に応じてバッファメモリ１０６から記憶媒体１０１へのデータの書き込み処理を行う（ステップＳ３４６）。

【０１５９】

そして、処理シーケンサ２０１は、ストレージ２１から受信準備完了のメッセージ（ステップＳ３４４において送った冗長化write コマンドに対する応答）が送信されるのを待つ（ステップＳ３４７）。ストレージ２１からの受信準備完了のメッセージを受信したことが通信部２０４から通知されると、バッファメモリ１０６に格納されたデータを冗長化するように冗長化部２０２に指示する。冗長化部２０２は、指示に応じて、第５の実施の形態の場合と同様に冗長化を行う（ステップＳ３４８）。ステップＳ３４８では、元データをＮ個に分割し、分割後のデータから冗長データを作成してもよい。あるいは、元データを分割せず、元データを複製したデータを冗長データとしてもよい。また、元データを分割した各データと同一のデータを複製し、複製した各データを冗長データとしてもよい。元データと同一のデータを複製し、元データおよび複製データをそれぞれ分割してもよい。

【０１６０】

続いて、処理シーケンサ２０１は、通信部２０４に、冗長化部２０２によって冗長化されたデータ群をストレージ２１に向けて送信するように指示する（ステップＳ３４９）。通信部２０４は、この指示に応じて、冗長化されたデータ群（分割後の元データおよび冗長データ）をストレージ２１に送信する。その後、ストレージ２１から受信完了のメッセージが送信されるのと、媒体制御部１０５からの書き込み完了通知とを待ち（ステップＳ３５０）、ストレージ２１からの受信完了のメッセージを受信したことが通信部２０４から通知され、かつ、媒体制御部１０５からの書き込み完了通知を受けると、ホスト１０に完了通知し（ステップＳ３５１）、処理を終了する。なお、ストレージ２１が、ストレージ２０から冗長化されたデータ群を受信するときの動作は、実施の形態５と同様である。

【０１６１】

図１７に示す場合と同様に、ストレージ２０とストレージ２１とがネットワーク１３，１４によって接続されていてもよい。そして、ステップＳ３４９において、通信部２０４が冗長化されたデータ群を送信するときには、元データと冗長データとをそれぞれ別々のネットワークを介して送信するようにしてもよい。

【０１６２】

第１の実施の形態から第４の実施の形態のデータ複製システムにおいて、ストレージと中継装置とがデータを送受信する際に、第５の実施の形態または第６の実施の形態に示したように冗長化されたデータ群を送受信するようにしてもよい。その場合、第１の実施の形態から第４の実施の形態においても、第５の実施の形態または第６の実施の形態と同様の効果が得られる。

【０１６３】

第５の実施の形態および第６の実施の形態において、データ転送処理手段は、処理シーケンサ２０１および通信部２０４によって実現される。冗長化手段２０２は、冗長株２０２によって実現される。復元手段は、復元部２０３によって実現される。格納処理手段は、処理シーケンサ２０１、ＩＯスケジューラ１０４および媒体制御部１０５によって実現される。

【０１６４】

実施の形態７.

図１９は、本発明によるデータ複製システムの第７の実施の形態を示すブロック図である。図１９に示すデータ複製システムにおいて、ストレージ３０１が、ストレージ３０１を使用するホスト（上位装置）３００とローカルに接続されている。また、ストレージ３０２が、ストレージ３０２を使用するホスト３０３とローカルに接続されている。ストレージ３０１は、ネットワーク１３を介してストレージ３０２に接続されている。また、ホスト３００とホスト３０３とは通信可能に接続されている。ホスト３００とホスト３０３とは、専用回線によって接続されていることが好ましいが、専用回線以外のネットワーク（例えばインターネット等）によって接続されていてもよい。

【０１６５】

ストレージ３０１，３０２は、例えば、単体の磁気ディスク装置、光ディスク装置または光磁気ディスク装置である。ストレージ３０１，３０２として、単体の磁気ディスク装置、光ディスク装置または光磁気ディスク装置の集合であるディスクアレイ装置を使用することもできる。ホスト３００，３０３とストレージ３０１，３０２とは、ＳＣＳＩ、ファイバチャネル（Fibre channel）、イーサ

ネット（登録商標）等で接続される。なお、図１９に示すシステムにおいて、ホスト３００が、システムに障害が発生していないときに稼働する正常系ホストであり、ホスト３０３が、ホスト３００において障害が発生したときに稼働する待機系ホストであるとする。

【０１６６】

ホスト３００，３０３は、それぞれホスト３００，３０３自身が保持するアプリケーションプログラム（以下、アプリケーションと記す。）に従って処理を行う。このアプリケーションは、ストレージ３０１，３０２のデータを処理対象とする。例えば、ストレージ３０１，３０２に銀行の顧客の預金額等のデータを記憶する場合には、ホスト３００，３０３は、預金データ管理アプリケーションに従って、ストレージ３０１，３０２内のデータの更新等を行う。実際に動作を行うのはホストであるが、以下、アプリケーションの動作として説明する。

【０１６７】

本実施の形態では、正常系のストレージ３０１内のデータが遠隔地に存在する待機系のストレージ３０３にミラーリングされる。ただし、正常系ホスト３００がストレージ３０１にデータを書き込んだ時点では、待機系ストレージ３０３の記憶媒体にそのデータを書き込ませるのではなく、待機系ストレージ３０３が備える同期用バッファメモリにそのデータを保持させる。そして、ホスト３００が指定するタイミングで、待機系ストレージ３０３の同期用バッファメモリ内のデータを所定の記憶媒体に書き込ませる。例えば、ホストが処理Ｘを開始するときには、ストレージにデータＡ，Ｂが書き込まれていなければならないとする。この場合、正常系のホスト３００がストレージ３０１にデータＡを書き込んだ時点では、待機系ストレージ３０２にはデータＡを同期用バッファメモリに保持させるだけで、待機系ストレージ３０２の記憶媒体への書き込みは行わせない。ホスト３００がストレージ３０１にデータＢを書き込んだ後に、ホスト３００は待機系ストレージ３０２の記憶媒体への書き込みタイミングを指定し、そのタイミングで待機系ストレージ３０２は同期用バッファメモリ内のデータＡ，Ｂを記憶媒体に書き込む。

【０１６８】

そのデータの状態であればアプリケーションがそのまま動作を再開することができるタイミングを、再開可能ポイントと記す。すなわち、再開可能ポイントとは、書き込まれたデータの状態がアプリケーションによる処理を再開することができる状態になっているタイミングのことである。上記の例では、正常系のストレージ３０１にデータＡ，Ｂが書き込まれてから次のデータが書き込まれるまでの間が、処理Ｘを開始することができる再開可能ポイントとなる。

【０１６９】

図２０は、ホスト３００の構成例を示すブロック図である。ホスト３００において、単数あるいは複数のアプリケーションが動作する。ここでは、２つのアプリケーション３１０ａ，３１０ｂを例示する。アプリケーション３１０ａ，３１０ｂは、ＩＯ管理部３１１を用いて、ストレージ３０１中のデータにアクセスする。また、ＩＯ管理部３１１は、ストレージ３０１に再開可能ポイントを通知するための再開可能ポイント通知部３１２を有する。アプリケーション３１０ａ，３１０ｂに従い、再開可能ポイント通知部３１２は、再開可能ポイントにおいて、ストレージ３０１に再開可能ポイントであることを知らせる処理（再開可能ポイント通知処理）を行う。また、ホスト３００の状態を監視するホスト監視部３１３が備えられている。なお、ホスト３０３の構成は、ホスト３００の構成と同じである。

【０１７０】

本実施の形態におけるアプリケーション３１０ａ，３１０ｂは、再開機能を有するアプリケーションである。すなわち、ストレージの記憶媒体１０１のデータ記録状態が所定の状態になっていれば処理を再開できる機能を実現するアプリケーションである。

【０１７１】

図２１は、図１９に示すストレージ３０１の構成例を示すブロック図である。なお、ストレージ３０２の構成も、図２１に示すような構成である。図２１に示すように、ストレージ３０１は、ストレージコントローラ３２０とストレージ本体である記憶媒体１０１とを含む。ストレージコントローラ３２０は、ホスト３００および他のストレージと通信を行う通信部３２２、各処理のシーケンスを管

理する処理シーケンサ３２１、記憶媒体１０１に対する処理命令の順序制御を行うＩＯスケジューラ１０４、ＩＯスケジューラ１０４が発行する処理命令に従って記憶媒体１０１の動作を制御する媒体処理部１０５、ホスト３００から記憶媒体１０１へのデータおよび記憶媒体１０１からホスト３００へのデータを一時記憶するバッファメモリ１０６、および他のストレージから送られてきたデータを一時保存する同期用バッファメモリ３２２を含む。処理シーケンサ３２１は、例えば、プログラムに従って動作するＣＰＵで実現される。ＩＯスケジューラ１０４および媒体制御部１０５の動作は、第１の実施の形態におけるＩＯスケジューラ１０４および媒体制御部１０５の動作と同様である。

【０１７２】

なお、同期用バッファメモリ３２２として半導体メモリが使用される場合もあるし、磁気ディスク装置、光ディスク装置または光磁気ディスク装置等のより大容量の記憶装置が使用される場合もある。

【０１７３】

アプリケーション３１０ａ，３１０ｂがストレージ３０１からデータを読み出す際、ホスト３００は、ストレージ３０１にデータの読み出しを要求する。ストレージ３０１の処理シーケンサ３２１は、ＩＯスケジューラ１０４等によって、要求されたデータを記憶媒体１０１からバッファメモリ１０６にコピーする。そして、そのデータをホスト３００に送信する。

【０１７４】

次に、ホスト３００がストレージに３０１にデータを書き込むときの動作の概要について説明する。ホスト３００は、write コマンドをストレージ３０１に出力することによって、ストレージ３０１にデータを書き込ませる。また、一回または複数回write コマンドを出力した後に再開可能ポイントになったならば、再開可能ポイント通知コマンドをストレージ３０１に出力することによって、再開可能ポイントになったことを通知する。

【０１７５】

ストレージ３０１は、write コマンドを受信した場合、そのwrite コマンドに従って、記憶媒体１０１にデータを書き込む。また、ストレージ３０２に、遅延

write コマンドを出力して、そのデータを待機系ストレージ３０２の同期用バッファメモリ３２３に保持させる。遅延write コマンドは、記憶媒体１０１に書き込むべきデータを同期用バッファメモリ３２３に保持させ、後述する遅延データ反映コマンドが届いたときに記憶媒体１０１に書き込むことを指示するコマンドである。

【０１７６】

ストレージ３０１は、再開可能ポイント通知コマンドを受信した場合、遅延データ反映コマンド（遅延書き込み実行要求）をストレージ３０２に送信する。遅延データ反映コマンドは、同期用バッファメモリ内に記憶しているデータを記憶媒体１０１に書き込むことを指示するコマンドである。待機系のストレージ３０２は、遅延データ反映コマンドを受信したときに、同期用バッファメモリ３２３に記憶したデータを記憶媒体１０１に書き込む。このような動作によって、ストレージ３０２の記憶媒体１０１が常に処理を開始できる状態に保つ。

【０１７７】

ストレージ３０１は、遅延write コマンドおよび遅延データ反映コマンドに同期ＩＤおよび発行ＩＤを付加して送信する。図２２は、同期ＩＤおよび発行ＩＤの例を示す説明図である。ストレージ３０１は、ホスト３００から再開可能ポイント通知コマンドを受信する度に同期ＩＤを更新する。従って、ある再開可能ポイントから次の再開可能ポイントまでの間に出力した遅延write コマンドには、同一の同期ＩＤが付加される。ただし、ストレージ３０１は、再開可能ポイント通知コマンドを受信したときに発行する遅延データ反映コマンドに、更新する直前の同期ＩＤを付加する。その後に発行する遅延write コマンドに更新後の同期ＩＤを付加する。図２２に示す例では、遅延write コマンドに更新後の同期ＩＤである「２４」や「２５」が付加され、その直近に出力された遅延データ反映コマンドに更新前の同期ＩＤである「２３」や「２４」が付加されている。

【０１７８】

また、ストレージ３０１は、遅延write コマンドまたは遅延データ反映コマンドを発行する度に（すなわち、遅延write コマンドまたは遅延データ反映コマンドを作成して送信する度に）発行ＩＤを更新する。図２２に示す例では、遅延wr

ite コマンドまたは遅延データ反映コマンドの発行ＩＤが１増加する場合の例を示している。

【０１７９】

ストレージ３０１は、ホスト３００からのwrite コマンドを受信すると、ストレージ３０２に遅延write コマンドを送信し、その後、遅延write コマンドに対応するデータを送信する。このデータの送信順序は、遅延write コマンドの発行順序と同一でなくてもよい。例えば、ストレージ３０１が、データｐ，ｑを書き込ませるために遅延write コマンドを発行し、その後データｒを書き込ませるために次の遅延write コマンドを発行するものとする。この場合、ストレージ３０１が各データを送信する順番は、ｐ，ｑ，ｒという順番に限られない。データｐ，ｑの送信完了前にデータｒの送信を開始してもよい。ただし、遅延データ反映コマンドをストレージ３０２に送信するときには、それまでに発行した遅延write コマンドのデータの送信を完了させてから遅延データ反映コマンドを送信する。

【０１８０】

なお、ストレージ３０１は、遅延write コマンドに対して、遅延write コマンドで指定されたデータを書き込むべき場所（オフセットアドレス、セクター番号、ブロック番号等）やデータのサイズも付加する。

【０１８１】

次に、アプリケーション３１０ａ，３１０ｂが、ストレージ３０１にデータを書き込むときの動作を説明する。図２３は、ストレージコントローラ３２０における処理シーケンサ３２１の動作を示すフローチャートである。

【０１８２】

アプリケーション３１０ａ，３１０ｂは、ストレージ３０１にデータを書き込むときに、ストレージ３０１のストレージコントローラ３２０における通信部３２２に対してwrite コマンドを出力する。通信部３２２は、write コマンドを受け取ると、write コマンドを処理シーケンサ３２１に渡し、write 処理を開始することを処理シーケンサ３２１に指示する。

【０１８３】

処理シーケンサ３２１は、write コマンドを受け取ると、図２３に示すように、ホスト３００から送られてくるデータを格納するのに必要な領域をバッファメモリ１０６に確保する（ステップＳ４００）。そして、準備完了の通知をホスト３００に送るように通信部３２２に指示する（ステップＳ４０１）。通信部３２２は、指示に応じて、ホスト３００に準備完了の通知を送信する。

【０１８４】

次いで、ホスト３００からデータが届くのを待ち（ステップＳ４０２）、データが届いて通信部３２２からデータを格納すべきバッファメモリ１０６の領域の問い合わせを受けると、ステップＳ４００で確保した領域を通信部３２２に知らせる（ステップＳ４０３）。また、通信部３２２にストレージ３０２に対して同期ＩＤを指定して遅延write コマンドを送信するよう指示する（ステップＳ４０４）。通信部３２２は、指示に応じて、ストレージ３０２に遅延write コマンドを送信する。既に説明したように、遅延write コマンドには、同期ＩＤおよび発行ＩＤが付加される。ある再開可能ポイントから次の再開可能ポイントまでの間に出力する各遅延write コマンドには、同一の同期ＩＤを付加する。また、遅延write コマンドまたは遅延データ反映コマンドを出力する度に発行ＩＤを更新する（図２２参照）。

【０１８５】

そして、ホスト３００からのデータのバッファメモリ１０６への格納の完了を待ち（ステップＳ４０５）、全てのデータがバッファメモリ１０６に格納されたことが通信部３２２から通知されると、処理シーケンサ３２１は、ＩＯスケジューラ１０４に対して、処理の種類（この場合には書き込み）と、処理の識別ＩＤと、処理の対象となる記憶媒体１０１中の領域を示す情報と、処理の対象となるバッファメモリ１０６の領域を示す情報とを登録する。ＩＯスケジューラ１０４は、書き込み要求に応じて、媒体制御部１０５に、書き込み対象のデータの書き込み指示を行う。媒体制御部１０５は、登録内容に応じてバッファメモリ１０６から記憶媒体１０１へのデータの書き込み処理を行う（ステップＳ４０６）。

【０１８６】

そして、ストレージ３０２から準備完了のメッセージ（遅延write コマンドに

対する応答）が送信されるのを待つ（ステップＳ４０７）。ストレージ３０２からの準備完了のメッセージを受信したことが通信部３２２から通知されると、処理シーケンサ３２１は、通信部３２２に、バッファメモリ１０６に格納されたデータをストレージ３０２に送信させる（ステップＳ４０８）。その後、ストレージ３０２から受信完了のメッセージが送信されるのと、媒体制御部１０５からの書き込み完了通知とを待ち（ステップＳ４０９）、ストレージ３０２からの受信完了のメッセージを受信したことが通信部３２２から通知され、かつ、媒体制御部１０５からの書き込み完了通知を受けると、ホスト３００に完了通知し（ステップＳ４１０）、処理を終了する。

【０１８７】

ステップＳ４０８において送信するデータの順番は、対応する遅延write コマンドの発行順序と異なっていてもよい。ストレージ３０１がホスト３００から連続してwrite コマンドを受信し、連続して遅延write コマンドを発行したとする。この場合、ストレージ３０１は、先に発行した遅延write コマンドに対応するデータの送信（ステップＳ４０８）が完了しないうちに、後の遅延write コマンドに対応するデータの送信（ステップＳ４０８）を開始してもよい。

【０１８８】

次に、ストレージ３０２が遅延write コマンドを受信したときの処理を説明する。図２４は、ストレージ３０２のストレージコントローラ３２０における処理シーケンサ３２１が遅延write コマンドを受信したときの動作を示すフローチャートである。

【０１８９】

処理シーケンサ３２１は、まず、受け取るデータに必要な領域を同期用バッファメモリ３２３に確保し（ステップＳ４２０）、要求の送信元（この例ではストレージ３０１）に受信準備完了の通知を送るように通信部３２２に指示する。通信部３２２は、指示に応じて受信準備完了のメッセージを送信元に送る（ステップＳ４２１）。次いで、データの到着を待ち（ステップＳ４２２）、データが到着し通信部３２２からデータの格納位置の問い合わせを受けると、ステップＳ４２０で確保した同期用バッファメモリ３２３の領域を通信部３２２に通知する（

ステップＳ４２３）。そして、要求の送信先からのデータの同期用バッファメモリ３２３への格納が完了するのを待ち（ステップＳ４２４）、データの格納が完了したら、要求の送信先に遅延write コマンドにもとづく処理を完了したことを通知するように通信部３２２に指示する。通信部３２２は、指示に応じて、遅延write コマンドにもとづく処理を完了したことを要求の送信先に通知する（ステップＳ４２５）。そして、処理シーケンサ３２１は処理を終了する。

【０１９０】

また、処理シーケンサ３２１は、データを記憶媒体１０１に書き込む際に、同期用バッファメモリ３２３に既に格納されているデータと同一の同期ＩＤを有し、かつ、格納する領域が一部だけでも重なるデータについては、既に同期用バッファメモリ３２３に格納されている重なる部分を書き込まないように制御する。すなわち、同じ領域に書き込むデータに関して、後続の遅延write コマンドにもとづく処理でのデータが有効になるように制御を行う。

【０１９１】

例えば、記憶媒体１０１中のある領域へのデータの書き込みを指示する遅延write コマンドがストレージ３０２に届き、同じ領域への書き込みを指示する遅延write コマンドが届いたとする。そして、この二つの遅延write コマンドの同期ＩＤが同一であるとする。この場合、正常系のストレージ３０１において再開可能ポイントの間に、一回データが書き込まれ、さらにデータが上書きされたことを意味する。従って、処理シーケンサ３２１は、同じ領域に書き込まれるデータのうち最初に書き込まれるデータを待機系ストレージ３０２の同期用バッファメモリ３２３の中に保持しなくてもよい。処理シーケンサ３２１は、新たにストレージ３０１から遅延write コマンドを受信したときに、上書きされることになるデータが同期用バッファメモリ３２３に存在するか否かを判断する。そして、上書きされることになるデータを特定したならばそのデータを同期用バッファメモリ３２３から削除する。

【０１９２】

次に、アプリケーション３１０ａ，３１０ｂが、ストレージ３０１に再開可能ポイント通知を通知する再開可能ポイント通知処理の動作を説明する。ここでは

、アプリケーション３１０ａが再開可能ポイント通知処理を行う場合を例にする。

【０１９３】

アプリケーション３１０ａが、再開可能ポイントを通知する場合、ＩＯ管理部３１１に再開可能ポイントを通知するように指示する。ＩＯ管理部３１１において、アプリケーション３１０ａから再開可能ポイントの通知が指示されると、再開可能ポイント通知部３１２が、ストレージ３０１に対して再開可能ポイント通知コマンドを発行する。

【０１９４】

ストレージ３０１に再開可能ポイント通知コマンドが到着すると、ストレージ３０１のストレージコントローラ３２０における通信部３２２に再開可能ポイント通知コマンドが入力される。通信部３２２は、再開可能ポイント通知コマンドを受け取ると、処理シーケンサ３２１に再開可能ポイント通知コマンドを渡し、再開可能ポイント通知処理の開始を指示する。

【０１９５】

図２５は、処理シーケンサ３２１の動作を示すフローチャートである。処理シーケンサ３２１は、まず、その時点での内部の同期ＩＤの値を保持した後に、同期ＩＤの値を更新する（ステップＳ４４０）。同期ＩＤの値を保持するとは、例えば、レジスタに保存することである。次いで、ステップＳ４４０で保持した同期ＩＤを指定した遅延データ反映コマンドをストレージ３０２に送信するように通信部３２２に指示する。通信部３２２は、指示に応じて、遅延データ反映コマンドをストレージ３０２に送信する（ステップＳ４４１）。そして、ストレージ３０２から遅延データ反映コマンドにもとづく処理の完了のメッセージが送られてくるのを待つ（ステップＳ４４２）。ストレージ３０２から完了のメッセージが到着したことが通信部３２２から通されたら、通信部３２２を介してホスト３００に再開可能ポイント通知コマンドにもとづく処理の完了を通知し（ステップＳ４４３）、処理を終了する。

【０１９６】

なお、図２２に示すように、一つの遅延データ反映コマンドの送信後に出力す

る各遅延write コマンドには、レジスタ等に保存した同期ＩＤではなく、更新後の同期ＩＤを付加する。

【０１９７】

次に、ストレージ３０２の遅延データ反映コマンドを受けた際の動作を説明する。ストレージ３０２に遅延データ反映コマンドが到着すると、ストレージ３０２のストレージコントローラ３２０の通信部３２２に遅延データ反映コマンドが入力される。通信部３２２は、遅延データ反映コマンドを受け取ると、処理シーケンサ３２１に遅延データ反映コマンドを渡し、遅延データ反映処理の開始を指示する。

【０１９８】

図２６は、ストレージ３０２のストレージコントローラ３２０における処理シーケンサ３２１の遅延データ反映処理の動作を示すフローチャートである。処理シーケンサ３２１は、到着した遅延データ反映コマンドに付加された同期ＩＤの値が前回処理した遅延データ反映コマンドの次の値かどうかを判定する（ステップＳ４６０）。次の値でなかった場合には、同期ＩＤが、前回処理した遅延データ反映コマンドの同期ＩＤの次の値になっている遅延データ反映コマンドの到着を待つ。例えば、図２２に示す例において、前回処理した遅延データ反映コマンドが、同期ＩＤ「２３」の遅延データ反映コマンドであったとする。その後、同期ＩＤ「２５」の遅延データ反映コマンドを受信した場合、同期ＩＤ「２４」の遅延データ反映コマンドの到着を待つ。そして、同期ＩＤ「２４」の遅延データ反映コマンドについてステップＳ４６１～Ｓ４６８の処理を行った後、同期ＩＤ「２５」の遅延データ反映コマンドについてステップＳ４６１以降の処理を行う。

【０１９９】

同期ＩＤが前回の同期ＩＤの次の値であった場合には、保持してある前回処理した遅延データ反映コマンドの発行ＩＤの値と、今回処理する遅延データ反映コマンドに付加された発行ＩＤの値の間の値の発行ＩＤを持つ遅延write コマンドに対応するデータが全て同期用バッファメモリ３２３に記録されているか否か検索する（ステップＳ４６１）。例えば、図２２に示す各コマンドのうち、前回処

理した遅延データ反映コマンドの発行ＩＤが「７１」であり、今回処理する遅延データ反映コマンドに付加された発行ＩＤが「７６」であるとする。この場合、発行ＩＤが「７１」～「７５」の遅延write コマンドに対応するデータが全て同期用バッファメモリ３２３に記録されているか否かを確認する。ただし、上書きされるデータであるとして削除したデータは記録確認の対象に含めなくてよい。抜けがあった場合には、各遅延write コマンドに対応するデータが全て待機系ストレージ３０２に到着し、同期用バッファメモリ３２３に記録されるまで待つ。

【０２００】

ステップＳ４６３では、処理シーケンサ３２１は、同期用バッファメモリ３２３を検索し、遅延データ反映コマンドにより指定された同期ＩＤと一致し、かつ、書き込み処理中でない遅延write コマンドを検索する。例えば、図２２に示す発行ＩＤ「７６」の遅延データ反映コマンドを受信して処理を行っている場合、そのコマンドに付加された同期ＩＤ「２４」と一致し、かつ、書き込み処理中でない遅延write コマンドを検索する。検索した結果、見つからなければ、記憶媒体１０１へのデータの書き込み処理が開始されていない遅延write コマンドが存在しないことになる。この場合、ステップＳ４６５に移行する。また、検索対象の遅延write コマンドが見つかった場合には、記憶媒体１０１へのデータの書き込み処理が開始されていない遅延write コマンドが存在することになる。この場合、ステップＳ４６４に移行する。

【０２０１】

ステップＳ４６４では、ステップＳ４６２の検索で見つかった遅延write コマンドで指定された記憶媒体１０１上のデータを書き込むべき場所（オフセットアドレス）に対して、同期用バッファメモリ３２３中の対応したデータを書き込む指示（書き込み要求）をＩＯスケジューラ１０４に登録する。ＩＯスケジューラ１０４は、書き込み要求に応じて、媒体制御部１０５に、データ転送対象のデータの書き込み指示を行う。媒体制御部１０５は、書き込み指示に従って、データを同期用バッファメモリ３２３から記憶媒体１０１に出力させる。そして、処理シーケンサ３２１は、同期用バッファメモリ３２３中の検索された領域を書き込み処理中とし、ステップＳ４６２に戻る。

【０２０２】

ステップＳ４６５では、同期用バッファメモリ３２３から記憶媒体１０１へデータ出力処理を行っているか否かを判断する。そして、処理中であればステップＳ４６６に移行し、すでに処理が完了しているならばステップＳ４６８に移行する。ステップＳ４６６では、記憶媒体１０１へデータ出力処理のうち１つが終了するまで待ち、完了した処理の対象であった同期用バッファメモリ３２３中の領域を未使用状態にして（ステップＳ４６７）、ステップＳ４６５に戻る。ステップＳ４６８では、遅延データ反映コマンドの送信元（この例ではストレージ３０１）に遅延データ反映コマンドにもとづく処理の完了を通知するように通信部３２２に指示する。通信部３２２は、指示に応じて、処理の完了を遅延データ反映コマンドの送信元に通知する。また、処理シーケンサ３２１は、前回処理した遅延データ反映コマンドの情報として保持している同期ＩＤおよび発行ＩＤを、処理した遅延データ反映コマンドのものに更新し、処理を終了する。

【０２０３】

次に、災害発生時の動作を説明する。災害等によってホスト３００が使用できなくなった場合には、ストレージ３０２を待機系から正常系にする。また、ホスト３０３が、ホスト３００から処理を引き継ぐ。

【０２０４】

障害検知からアプリケーション再開までのホスト３０３の動作を説明する。まず、ホスト３０３中のホスト監視部３１３がホスト３００の異常を検出する。ホスト監視部３１３は、ホスト３００の異常を検出すると、待機系であるストレージ３０２に対して遅延データ破棄コマンドを発行する。そして、遅延データ破棄コマンドに対する応答を待ち、応答を受けたら、待機系ホスト３０３のホスト監視部３１３は、待機系ホストのアプリケーション３１０ａ，３１０ｂを実行させる。以後、ホスト３０３からのデータの書き込みおよび読み出しは、ストレージ３０２に対して行われる。

【０２０５】

なお、ホスト３０３中のホスト監視部３１３は、ホスト３００と常時あるいは定期的に通信を行っている。ホスト監視部３１３が、一定時間ホスト３００と通

信できなくなった場合、あるいは、ホスト３００から異常が報告された場合に、ホスト監視部３１３は、ホスト３００において災害が発生したと認識する。ホスト３０３がホスト３００の異常を確実に認識するために、ホスト３０３とホスト３００とは専用回線で接続されていることが好ましい。

【０２０６】

次に、待機系ストレージ３０２が、待機系ホスト３０３から遅延データ破棄コマンドを受けた場合の動作を説明する。ストレージ３０２に遅延データ破棄コマンドが到着すると、ストレージ３０２のストレージコントローラ３２０における通信部３２２に遅延データ破棄コマンドが入力される。通信部３２２は、遅延データ破棄コマンドを受け取ると、処理シーケンサ３２２に遅延データ破棄コマンドを渡し、遅延データ破棄処理の開始を指示する。

【０２０７】

図２７は、処理シーケンサ３２２が実行する遅延データ破棄処理を示すフローチャートである。処理シーケンサ３２２は、同期用バッファメモリ３２３を検索する（ステップＳ４８０）。ステップＳ４８０では、未だ遅延データ反映コマンドが到着していないため記憶媒体１０１に移動させる必要がないデータを検索する。処理シーケンサ３２２は、検索した結果、見つからなかった場合ステップＳ４８３に移行し、見つかった場合にはステップＳ４８２に移行する（ステップＳ４８１）。例えば、図２２に示す各コマンドのうち、発行ＩＤ「７６」の遅延データ反映コマンドを受信しておらず、発行ＩＤ「７２」～「７５」のうちの一部の遅延write コマンドのデータが同期用バッファメモリ３２３に記憶されているならば、ステップＳ４８２に移行する。発行ＩＤ「７６」の遅延データ反映コマンドを受信して、各遅延write コマンドのデータを記憶媒体１０１に移行している場合には、ステップＳ４８３に移行する。

【０２０８】

ステップＳ４８２では、処理シーケンサ３２２は、ステップＳ４８０で見つかったデータを記録している同期用バッファメモリ３２３中の領域を未使用状態にしてステップＳ４８０に戻る。すると、次のステップＳ４８１では、ステップＳ４８３に移行することになる。ステップＳ４８３では、遅延データ反映処理中で

あった場合にはステップＳ４８４に移行し、遅延データ反映処理中でなかった場合にはステップＳ４８５に移行する。

【０２０９】

ステップＳ４８４では、処理中の遅延データ反映処理の完了を持ち、完了したらステップＳ４８５に移行する。ステップＳ４８５では、遅延データ破棄コマンドの発行元（この例ではホスト３０３）に遅延データ破棄コマンドにもとづく処理の完了を通知するように通信部３２２に指示する。通信部３２２は、指示に応じて、処理の完了を遅延データ破棄コマンドの発行元に通知する、そして、処理シーケンサ３２２は処理を終了する。

【０２１０】

本実施の形態では、正常系のホスト３００がストレージ３０１にwrite コマンドを出力したとき、待機系のストレージ３０２はストレージ３０１から書き込まれるデータを受信するが、記憶媒体１０１には記録せず、同期用バッファメモリ３２３に記録する。そして、ストレージ３０２は、再開可能ポイント通知コマンドを受信したときに、そのデータを同期用バッファメモリ３２３から記憶媒体１０１に移動させ、遅延データ破棄コマンドを受信した場合、同期用バッファメモリ３２３を未使用状態にする。従って、ストレージ３０２の記憶媒体１０１は、常にアプリケーション３１０ａ，３１０ｂの処理を再開できる状態に保たれる。その結果、ホスト３００に異常が生じたときに、即座にホスト３０３が処理を続行することができる。

【０２１１】

また、正常系ストレージ３０１は、再開可能ポイントと再開可能ポイントとの間で記憶媒体１０１に書き込んだデータを一度に待機系ストレージ３０２に送信するのではなく、write コマンドを受信したタイミング毎に送信する。従って、大量のデータを一度に送信しないので、ホスト３０２とのデータ転送時間が少なくてすむ。

【０２１２】

さらに、待機系ストレージ３０２がストレージ３０１から受信する各コマンドは同期ＩＤおよび発行ＩＤによって管理され、再開可能ポイントと再開可能ポイ

ントとの間に正常系ストレージが送信したデータの到着順序は任意の順序でよい。従って、遅延write コマンドに対応するデータを遅延write コマンドが送信する順序が制限されないので、設計を行いやすくなる。

【０２１３】

本実施の形態では、遅延write コマンドおよび遅延データ反映コマンドに同期ＩＤおよび発行ＩＤの双方を付加する場合を説明したが、発行ＩＤのみを付加するようにしてもよい。この場合、待機系ストレージ３０２の処理シーケンサ３２１は、遅延データ反映コマンドを受信した場合、ステップＳ４６０，Ｓ４６１の代わりに以下の処理を行えばよい。処理シーケンサ３２１は、遅延データ反映コマンドを受信した場合、前回処理した遅延データ反映コマンドの発行ＩＤと受信した遅延データ反映コマンドの発行ＩＤとの間の発行ＩＤが付加された各コマンドを全て受信しているか否かを確認する。そして、各発行ＩＤが付加されたコマンドを全て受信してれば、ステップＳ４６２（図２６参照）以降の処理を行う。各発行ＩＤが付加されたコマンドを全て受信していなければ、各コマンドを全て受信したときにステップＳ４６２以降の処理を開始する。

【０２１４】

例えば、図２２に示す各コマンド（同期ＩＤは付加されていないものとする）のうち、発行ＩＤ「７１」の遅延データ反映コマンドが前回処理した遅延データ反映コマンドであったとする。その後、発行ＩＤ「７６」の遅延データ反映コマンドを受信した場合、発行ＩＤ「７２」～「７５」の各コマンドを受信しているか否かを確認し、これらのコマンドを全て受信した後にステップＳ４６２以降の処理を開始する。また、発行ＩＤ「７６」より先に発行ＩＤ「８０」の遅延データ反映コマンドを受信した場合、制御シーケンサ３２１は、発行ＩＤ「７２」～「７９」の各コマンドを受信しているか否かを確認する。そして、発行ＩＤ「７２」～「７６」の各コマンドを全て受信したときに、発行ＩＤ「７６」の遅延データ反映コマンドの処理を行う。その後、発行ＩＤ「７７」～「７９」のコマンドを全て受信したときに発行ＩＤ「８０」の遅延データ反映コマンドの処理を行う。

【０２１５】

なお、各コマンドに同期ＩＤも付加される場合には、遅延write コマンドを受信したときに、上書きされるデータを特定することができる。一方、同期ＩＤを用いない場合には、処理シーケンサ３２１はステップＳ４６２の処理を開始する直前に、上書きされるデータを特定し、そのデータを同期用バッファメモリ３２３から削除する。

【０２１６】

第７の実施の形態において、遅延書き込み要求手段および書き込み実行要求手段は、ストレージ３０１の処理シーケンサ３２１および通信部３２２によって実現される。一時記憶手段は、同期用バッファメモリ３２３によって実現される。格納処理手段は、ストレージ３０２の処理シーケンサ３２１、ＩＯスケジューラ１０４および媒体制御部１０５によって実現される。再開可能ポイント通知手段は、再開可能ポイント通知部３１２によって実現される。

【０２１７】

実施の形態８.

図２８は、本発明によるデータ複製システムの第８の実施の形態を示すブロック図である。この実施の形態では、ストレージが主体となってデータ転送を実行するとともに、正常系システムにおけるストレージ内のデータが遠隔地にミラーリングされる。さらに、システムには、少なくとも１世代前のスナップショットを保存できる待機系のストレージが設けられる。１世代前のスナップショットとは、直近の再開可能ポイントにおいてデータが格納されていた記憶媒体１０１のアドレスの情報である。

【０２１８】

図２８に示すデータ複製システムにおいて、ストレージ４００が、ストレージ４００を使用するホスト３００とローカルに接続されている。また、ストレージ４０１が、ストレージ４０１を使用するホスト３０３とローカルに接続されている。ストレージ３０１は、ネットワーク１３を介してストレージ３０２に接続されている。また、ホスト３００とホスト３０３とは、ホスト３０３がホスト３００の状態を監視するために通信可能に接続されている。ホスト３００とホスト３０３は専用回線によって接続されていることが好ましいが、専用回線以外のネッ

トワーク（例えばインターネット等）によって接続されていてもよい。なお、ホスト３００，３０３の構成は、第７の実施の形態におけるホスト３００，３０３の構成と同じである（図２０参照）。

【０２１９】

ストレージ４００，４０１は、例えば、単体の磁気ディスク装置、光ディスク装置または光磁気ディスク装置である。ストレージ４００，４０１として、単体の磁気ディスク装置、光ディスク装置または光磁気ディスク装置の集合であるディスクアレイ装置を使用することもできる。ホスト３００，３０３とストレージ４００，４０１とは、ＳＣＳＩ、ファイバチャネル（Fibre channel ）、イーサネット（登録商標）等で接続される。なお、図２８に示すシステムにおいて、ホスト３００が、システムに障害が発生していないときに稼働する正常系ホストであり、ホスト３０３が、ホスト３００において障害が発生したときに稼働する待機系ホストであるとする。

【０２２０】

この実施の形態では、ホスト３００上で動作するアプリケーション３１０ａ，３１０ｂが、動作中に、そのデータの状態であればアプリケーションがそのまま動作を再開可能なポイントで、ストレージ３０１に再開可能なポイントであることを知らせるために再開可能ポイント通知処理を行う。なお、アプリケーション３１０ａ，３１０ｂが、ストレージ４００からデータを読み出す際のホスト３００およびストレージ４００の動作は、通常のデータ読み出し処理の場合の動作と同様である。また、ストレージ４００において、ＩＯスケジューラ１０４および媒体制御部１０５の動作は、第１の実施の形態におけるＩＯスケジューラ１０４および媒体制御部１０５の動作と同様である。

【０２２１】

本実施の形態におけるアプリケーション３１０ａ，３１０ｂは、再開機能を有するアプリケーションである。すなわち、ストレージの記憶媒体１０１のデータ記録状態が所定の状態になっていれば処理を再開できるアプリケーションである。

【０２２２】

図２９は、図２８に示すストレージ４００の構成例を示すブロック図である。なお、ストレージ４０１の構成も、図２９に示すような構成である。図２９に示すように、ストレージ４００は、ストレージコントローラ４１０とストレージ本体である記憶媒体１０１とを含む。ストレージコントローラ４１０は、ホスト３００および他のストレージと通信を行う通信部４１２、各処理のシーケンスを管理する処理シーケンサ４１１、記憶媒体１０１に対する処理命令の順序制御を行うＩＯスケジューラ１０４、ＩＯスケジューラ１０４が発行する処理命令に従って記憶媒体１０１の動作を制御する媒体処理部１０５、ホスト３００から記憶媒体１０１へのデータおよび記憶媒体１０１からホスト３００へのデータを一時記憶するバッファメモリ１０６、論理ブロックアドレスの管理を行うＬＢＡ管理部４１３、および論理ブロックアドレスの管理に用いられるアドレステーブルを記憶したアドレステーブル記憶部４１４を含む。処理シーケンサ３２１は、例えば、プログラムに従って動作するＣＰＵで実現される。

【０２２３】

なお、同期用バッファメモリ３２２として半導体メモリが使用される場合もあるし、磁気ディスク装置、光ディスク装置または光磁気ディスク装置等のより大容量の記憶装置が使用される場合もある。

【０２２４】

ストレージ４００，４０１は、記憶媒体１０１の物理アドレス空間によって、データを管理する。一方、他の装置からデータの書き込み処理や読み込み処理等の要求を受け付ける場合には、書き込み領域や読み込み領域を論理アドレス空間のアドレスによって指定される。再開可能ポイントにおける論理アドレス空間をスナップショット論理アドレス空間と呼ぶ。また、最新の時点において論理アドレスを最新論理アドレス空間と呼ぶ。ストレージ４００，４０１は、スナップショット論理アドレス空間の情報と、最新論理アドレス空間の情報とを用いて処理を進める。また、ストレージ４００，４０１は、複数のスナップショット論理アドレス空間を管理する場合もある。

【０２２５】

アドレステーブル記憶部４１４が記憶するアドレステーブルの一例を図３０に

示す。アドレステーブル４２０は、複数のエントリ４２０－０～４２０－ｎから構成されている。各エントリ４２０－０～４２０－ｎは、論理アドレス空間を固定長のデータに分割した論理ブロックに対応する。

【０２２６】

図３１（ａ）は、エントリの説明図である。各エントリ４２０－０～４２０－ｎを構成する情報には、最新論理アドレス空間に対応したストレージ内の物理ブロック番号と、スナップショット論理アドレス空間に対応した物理ブロック番号がある。各物理ブロック番号は、直近にスナップショットを作成してからブロック番号の変更があったことを記録するフラグとセットになっている。なお、物理ブロック番号とは、ストレージ４００，４０１のアドレス空間を固定長のデータに分割した際の物理ブロック個々につけられた一意の番号である。なお、スナップショット論理アドレス空間に対応する物理ブロック番号のフラグは設けなくてもよい。

【０２２７】

図３１（ｂ）は、一つのエントリの初期状態の例を示す。図３１（ｂ）では、「ａａａ」という物理ブロック番号に対応するアドレスにデータが格納されていることを示す。また、再開可能ポイント以降、データが変更されていないので、最新論理アドレス空間に対応する物理ブロック番号も「ａａａ」となっている。「ａａａ」という物理ブロック番号に対応するアドレスへの書き込みが要求されたとする。その場合、書き込みは他のアドレスに行い、そのアドレスに対応する物理ブロック番号「ｂｂｂ」を最新論理アドレス空間に対応する物理ブロック番号として保持する。また、新論理アドレス空間に対応するフラグも「変更有り」に更新する。その後、再開可能ポイントになったならば、図３１（ｄ）に示すように最新論理アドレス空間に対応する情報を、スナップショット論理アドレス空間に対応する情報にコピーする。また、最新論理アドレス空間に対応するフラグを「変更無し」とする。

【０２２８】

本実施の形態では、スナップショット論理アドレス空間に対応する情報をスナップショットとして用いる。

【０２２９】

この結果、次に再開可能ポイントとなるまでは、新たに書き込もうとしたデータは、書き込もうとしたアドレスとは別のアドレスに書き込まれる。従って、再開可能ポイントの時点におけるデータが残ることになる。待機系のストレージへのデータの書き込みをこのように管理することにより、ホスト３００に異常が生じても、すぐにホスト３０３が処理を再開することができる。

【０２３０】

また、アドレステーブル記憶部４１４は、どの論理アドレス空間にも含まれない物理ブロックのブロック番号が記録された未使用ブロックテーブルも記憶する。なお、アドレステーブル記憶部４１４として、不揮発性の半導体メモリ、磁気ディスク装置、光ディスク装置または光磁気ディスク等が使用される。また、アドレステーブル記憶部４１４として、記憶媒体１０１の記憶領域の一部が使用される場合もある。

【０２３１】

次に、ホスト３００が、ストレージ４００，４０１にデータを書き込むときの動作を説明する。ホスト３００は、ストレージ４００にデータを書き込むときに、ストレージ４００のストレージコントローラ４１０における通信部４１２に対してwrite コマンドを出力する。通信部４１２は、write コマンドを受け取ると、write コマンドを処理シーケンサ４１１に渡し、write 処理を開始することを処理シーケンサ４１１に指示する。処理シーケンサ４１１が実行する書き込み処理は、データ転送先ストレージが設定されている場合といない場合とで異なる。なお、ストレージ４００には、データ転送先としてストレージ４０１が設定され、ストレージ４０１には、データ転送先が設定されていない。

【０２３２】

図３２は、データ転送先が設定されている場合のストレージコントローラ４１０における処理シーケンサ４１１の動作を示すフローチャートである。すなわち、この実施の形態では、ストレージ４００における処理シーケンサ４１１の動作を示すフローチャートである。データ転送先が設定されている場合には、処理シーケンサ４１１は、まず、受け取るデータに必要な領域を同期用バッファメモリ

３２３に確保する（ステップＳ５００）。また、write コマンドで指定された書き込み先の論理アドレスとwrite 処理であることを指定して、ＬＢＡ管理部４１３に論理アドレスを物理アドレスに変換させる（ステップＳ５０１）。ＬＢＡ管理部４１３は、図３１に示す最新論理アドレス空間に対応する物理ブロック番号を割り当て、その物理ブロック番号に対応する物理アドレスを処理シーケンサ４１１に返す。ステップＳ５０１におけるＬＢＡ管理部４１３の動作の詳細については後述する。

【０２３３】

次いで、処理シーケンサ４１１は、準備完了の通知をホスト３００に送るように通信部４１２に指示する（ステップＳ５０２）。通信部４１２は、指示に応じて、ホスト３００に準備完了の通知を送信する。そして、ホスト３００からデータが届くのを待ち（ステップＳ５０３）、データが届いて通信部４１２からデータを格納すべきバッファメモリ１０６の領域の問い合わせを受けると、ステップＳ５００で確保した領域を通信部４１２に知らせる（ステップＳ５０４）。また、再開可能ポイント通知処理を行っている場合にはステップＳ５１３に移行し、行っていなかった場合にはステップＳ５０６に移行する（ステップＳ５０５）。ここで、「再開可能ポイント通知処理を行っている場合」とは、ホスト３００から再開可能ポイントを通知され、ストレージ４０１に対して所定の処理を行っている場合を指す。具体的には、後述するステップＳ５６２～Ｓ５６４の処理を行っている場合にはステップＳ５１３に移行し、行っていなかった場合にはステップＳ５０６に移行する。

【０２３４】

ステップＳ５０６では、処理シーケンサ４１１は、データ転送先のストレージ（この例ではストレージ４０１）に対してwrite コマンドを発行するように通信部４１２に指示する。このwrite コマンドでは、書き込み先を論理アドレスで指定する。通信部４１２は、指示に応じて、write コマンドをデータ転送先のストレージに送信する。そして、ホスト３００からのデータのバッファメモリ１０６への格納の完了を待ち（ステップＳ５０７）、全てのデータがバッファメモリ１０６に格納されたことが通信部４１２から通知されると、処理シーケンサ４１１

は、ＩＯスケジューラ１０４に対して、ステップＳ５０１で変換した物理アドレスに対応した領域に、バッファメモリ１０６に格納されたデータを書き込む指示（書き込み要求）を登録する。ＩＯスケジューラ１０４は、書き込み要求に応じて、媒体制御部１０５に、書き込み対象のデータの書き込み指示を行う。媒体制御部１０５は、登録内容に応じてバッファメモリ１０６から記憶媒体１０１へのデータの書き込み処理を行う（ステップＳ５０８）。

【０２３５】

そして、ストレージ４０１から準備完了のメッセージが送信されるのを待ち（ステップＳ５０９）、ストレージ４０１からの準備完了のメッセージを受信したことが通信部４１２から通知されると、通信部４１２に、バッファメモリ１０６に格納されたデータをストレージ４０１に送信させる（ステップＳ５１０）。その後、ストレージ４０１から受信完了のメッセージが送信されるのと、媒体制御部１０５からの書き込み完了通知とを待ち（ステップＳ５１１）、ストレージ４０１からの受信完了のメッセージを受信したことが通信部４１２から通知され、かつ、媒体制御部１０５からの書き込み完了通知を受けると、ホスト３００に完了通知し（ステップＳ５１２）、処理を終了する。

【０２３６】

ステップＳ５１３では、ホスト３００からのデータのバッファメモリ１０６への格納の完了を待ち、全てのデータがバッファメモリ１０６に格納されたことが通信部４１２から通知されると、処理シーケンサ４１１は、ＩＯスケジューラ１０４に対して、ステップＳ５０１で変換した物理アドレスに対応した領域に、バッファメモリ１０６に格納されたデータを書き込む指示（書き込み要求）を登録する（ステップＳ５１４）。次いで、再開可能ポイント通知処理が完了するまで待ち（ステップＳ５１５）、再開可能ポイント通知処理が完了したら、データ転送先のストレージに対してwrite コマンドを発行するように通信部４１２に指示する（ステップＳ５１６）。このwrite コマンドでは、書き込み先を論理アドレスで指定する。ステップＳ５１６の後、ステップＳ５０９に移行する。

【０２３７】

なお、ステップＳ５０６～Ｓ５０８での処理と、ステップＳ５１３～Ｓ５１６

での処理では、ストレージ４０１に対するwrite コマンド発行処理と記憶媒体１０１への書き込み処理の順番が逆になっている。これは、以下の理由によるものである。ステップＳ５０６～Ｓ５０８では、ストレージ４０１にwrite コマンドを送信し、その応答が戻って来るまでの期間に記憶媒体１０１への書き込み処理を進めるため、write コマンド発行処理を記憶媒体１０１への書き込み処理よりも先に行うこととした。一方、ステップＳ５１３～Ｓ５１６では、再開可能ポイント通知処理の終了を待っている間、ストレージ４０１にエントリの変更を行わせている。従って、再開可能ポイント通知処理の終了まではストレージ４０１にwrite コマンドを送信できない。そこで、再開可能ポイント通知処理の終了を待つまでの間に記憶媒体１０１への書き込み処理を進めることとした。

【０２３８】

図３３は、データ転送先が設定されている場合のストレージコントローラ４１０における処理シーケンサ４１１の動作を示すフローチャートである。すなわち、この実施の形態では、ストレージ４０１における処理シーケンサ４１１の動作を示すフローチャートである。より具体的には、ステップＳ５０６，Ｓ５１６で送信されたwrite コマンドを受信したストレージ４０１の処理シーケンサ４１１の動作を示すフローチャートである。

【０２３９】

処理シーケンサ４１１はwrite コマンドを受信すると、まず、受け取るデータに必要な領域をバッファメモリ１０６に確保する（ステップＳ５２０）。また、write コマンドで指定された書き込み先の論理アドレスとwrite 処理であることを指定して、ＬＢＡ管理部４１３に論理アドレスを物理アドレスに変換させる（ステップＳ５２１）。この処理はステップＳ５０１（図３２参照）と同様の処理である。

【０２４０】

次いで、処理シーケンサ４１１は、準備完了の通知をデータ転送元のストレージ（この例ではストレージ４００）に送るように通信部４１２に指示する（ステップＳ５２２）。通信部４１２は、指示に応じて、データ転送元のストレージに準備完了の通知を送信する。そして、データ転送元のストレージからデータが届

くのを待ち（ステップＳ５２３）、データが届いて通信部４１２からデータを格納すべきバッファメモリ１０６の領域の問い合わせを受けると、ステップＳ５２０で確保した領域を通信部４１２に知らせる（ステップＳ５２４）。

【０２４１】

さらに、データ転送元のストレージからのデータのバッファメモリ１０６への格納の完了を待ち（ステップＳ５２５）、全てのデータがバッファメモリ１０６に格納されたことが通信部４１２から通知されると、処理シーケンサ４１１は、ＩＯスケジューラ１０４に対して、ステップＳ５２１で変換した物理アドレスに対応した領域に、バッファメモリ１０６に格納されたデータを書き込む指示（書き込み要求）を登録する。ＩＯスケジューラ１０４は、書き込み要求に応じて、媒体制御部１０５に、書き込み対象のデータの書き込み指示を行う。媒体制御部１０５は、登録内容に応じてバッファメモリ１０６から記憶媒体１０１へのデータの書き込み処理を行う（ステップＳ５２６）。

【０２４２】

そして、媒体制御部１０５からの書き込み完了通知を待ち（ステップＳ５２７）、媒体制御部１０５からの書き込み完了通知を受けると、データ転送元のストレージに完了通知を行い（ステップＳ５２８）、処理を終了する。

【０２４３】

ストレージ４００，４０１は、いずれも論理アドレスを指定されたwrite コマンドを受信して、書き込み処理を開始する。しかし、その論理アドレスに対応する物理アドレスにデータを書き込むのではなく、別の物理アドレス（ステップＳ５０１，Ｓ５２１で割り当てた物理アドレス）にデータを書き込む。従って、再開可能ポイントにおけるデータを記憶媒体１０１に保持し続けることができるので、正常系のホスト３００に異常が生じても、待機系のホスト３０３は、短時間で処理を再開することができる。

【０２４４】

次に、ＬＢＡ管理部４１３による論理アドレスから物理アドレスへの変換処理を説明する。なお、ＬＢＡ管理部４１３による変換処理は、write 処理であることが指定された場合とされなかった場合とで異なる。write 処理であることが指

定されなかった場合には、ＬＢＡ管理部４１３は、論理アドレスが指定されて変換処理の指示を受けると、論理アドレスから論理ブロック番号を算出し、アドレステーブル記憶部４１４から、最新論理アドレス空間でのその論理ブロック番号に対応した物理ブロック番号を取得する。また、物理ブロック番号と論理ブロックアドレスとから物理アドレスを算出し、算出した物理アドレスを処理シーケンサ４１１に通知する。なお、論理ブロック番号は、論理ブロックアドレスをブロック長で割った値（小数点以下は切り捨てる）である。また、物理アドレスを算出する場合、まず物理ブロックとブロック長の積（Ｐとする）を求める。また、論理アドレスをブロック長で割った値の余り（Ｑとする）を求める。ＰとＱの和を物理アドレスとして算出する。

【０２４５】

図３４は、write 処理であることが指定された場合のＬＢＡ管理部４１３の動作を示すフローチャートである。すなわち、ステップＳ５０１，Ｓ５２１におけるＬＢＡ管理部４１３の動作を示すフローチャートである。write 処理であることが指定された場合には、ＬＢＡ管理部４１３は、論理アドレスから論理ブロック番号を算出する（ステップＳ５４０）、次いで、アドレステーブル記憶部４１４から、ステップＳ５４０で算出した論理ブロック番号に対応したエントリを取得する（ステップＳ５４１）。ステップＳ５４１では、最新論理アドレス空間におけるエントリを取得する。そして、エントリ中の、最新論理アドレス空間に対応する物理ブロック番号と対になるフラグを確認する（ステップＳ５４２）。フラグが「変更有り」となっている場合にはステップＳ５４６に移行し、「変更無し」となっている場合にはステップＳ５４３に移行する。

【０２４６】

フラグが「変更無し」となっている場合とは、図３１（ｂ）に示すように、最新論理アドレス空間に対応する物理ブロック番号がスナップショット論理アドレス空間に対応する物理ブロック番号と等しい場合である。以下、図３１（ｂ），（ｃ）を用いて説明する。ステップＳ５４３では、ＬＢＡ管理部４１３は、アドレステーブル記憶部４１４中の未使用ブロックテーブルから未使用の物理ブロック番号を入手する（ステップＳ５４３）。ここでは、「ｂｂｂ」という物理ブロ

ック番号を入手したものとする。また、アドレステーブル記憶部４１４に登録されている未使用ブロックテーブルの情報から、入手した物理ブロック番号の情報を削除する。そして、削除後の未使用ブロックテーブルの情報をアドレステーブル記憶部４１４に登録し直す。

【０２４７】

さらに、ステップＳ５４３で入手した物理ブロック番号（ｂｂｂ）に該当する領域に、ステップＳ５４１で入手したエントリ中の最新論理アドレス空間に対応する物理ブロック（ａａａ）に該当する領域のデータをコピーする旨の情報（コピー要求）をＩＯスケジューラ１０４に登録する。ＩＯスケジューラ１０４は、コピー要求に応じて、媒体制御部１０５に、対象のデータのコピー指示を行う。媒体制御部１０５は、記憶媒体１０１における指定された領域間のコピー処理を行う（ステップＳ５４４）。コピー処理が完了したら、ＬＢＡ管理部４１３は、ステップＳ５４１で入手したエントリの最新論理アドレス空間に対応する物理ブロック番号（ａａａ）をステップＳ５４３で入手した物理ブロック番号（ｂｂｂ）に変更し、変更後の物理ブロック番号とセットとなっているフラグを「変更有り」に更新する。また、アドレステーブル記憶部４１４に、変更したエントリを記憶させる（ステップＳ５４５）。そして、ステップＳ５４６に移行する。

【０２４８】

ステップＳ５４６では、ＬＢＡ管理部４１３は、エントリの最新論理アドレス空間に対応する物理ブロック番号から物理アドレスを算出し、算出した物理アドレスを処理シーケンサ４１１に通知し（ステップＳ５４７）、処理を終了する。

【０２４９】

ステップＳ５４２の後、ステップＳ５４３以降の処理を行うと、図３１（ｂ）に示す物理ブロック番号は、図（ｃ）に示すように変更される。しかし、ステップＳ５４７までの処理が終了した時点では、「ｂｂｂ」に対応する物理アドレスには、「ａａａ」に対応する物理アドレスと同一のデータが格納されている。正常系ストレージ４００が「ｂｂｂ」に対応する物理アドレスのデータを書き換えるのは、ステップＳ５０８またはステップＳ５１４においてである（図３２参照）。また、待機系ストレージ４０１が「ｂｂｂ」に対応する物理アドレスのデー

タを書き換えるのはステップＳ５２６においてである（図３３参照）。

【０２５０】

次に、アプリケーション３１０ａ，３１０ｂが、ストレージ４００に再開可能ポイント通知を通知するときの動作を説明する。アプリケーション３１０ａ，３１０ｂは、ストレージ４００に再開可能ポイント通知を通知するときに、ＩＯ管理部３１１（図２０参照）に再開可能ポイントの通知を指示する。すると、ＩＯ管理部３１１において再開可能ポイント通知部３１２は、ストレージ４００に対して再開可能ポイント通知コマンドを発行する。

【０２５１】

ストレージ４００に再開可能ポイント通知コマンドが到着すると、再開可能ポイント通知コマンドは、ストレージコントローラ４１０における通信部４１２に入力される。通信部４１２は、再開可能ポイント通知コマンドを受け取ると、再開可能ポイント通知コマンドを処理シーケンサ４１１に渡し、再開可能ポイント通知処理を開始することを処理シーケンサ４１１に指示する。

【０２５２】

図３５は、再開可能ポイント通知コマンドを受け取った処理シーケンサ４１１の動作を示すフローチャートである。処理シーケンサ４１１は、再開可能ポイント通知コマンドを受け取ると、図３５に示すように、まず、他ストレージに対して要求したwrite 処理のうちで完了していないものがあるか否かを調べ（ステップＳ５６０）、あった場合にはステップＳ５６１に処理を実行する。すなわち、他ストレージに対して要求したwrite 処理が全て完了するのを待つ（ステップＳ５６１）。なお、処理シーケンサ４１１は、発行した各write コマンドに対する応答が完了しているのか否かを示す一覧情報を管理する。ステップＳ５６０では、この情報に基づいて判断を行えばよい。

【０２５３】

処理シーケンサ４１１は、他ストレージに対して要求したwrite 処理が全て完了している状態で、通信部４１２に、データ転送先のストレージ（この例ではストレージ４０１）にスナップショット作成コマンド（スナップショット作成要求）を送信するように指示する。通信部４１２は、指示に応じて、データ転送先の

ストレージにスナップショット作成コマンドを送信する（ステップＳ５６２）。スナップショット作成コマンドとは、最新論理アドレス空間に対応する情報を、スナップショット論理アドレス空間に対応する情報にコピーするようにし、かつ、最新論理アドレス空間に対応するフラグを「変更無し」に更新するように要求するコマンドである。処理シーケンサ４１１は、ステップＳ５６２の後、転送先ストレージからの応答を待ち（ステップＳ５６３）、応答が到着ことが通信部４１２から通知されたら、通信部４１２にホスト３００に対して完了を通知するように指示し（ステップＳ５６４）、処理を終了する。

【０２５４】

次に、データ転送先のストレージであるストレージ４０１がスナップショット作成コマンドを受信した場合の動作を説明する。ストレージ４０１にスナップショット作成コマンドが到着すると、ストレージ４０１のストレージコントローラ４１０の通信部４１２にスナップショット作成コマンドが入力される。通信部４１２は、スナップショット作成コマンドを受け取ると、処理シーケンサ４１１にスナップショット作成コマンドを渡し、スナップショット作成処理の開始を指示する。

【０２５５】

スナップショット作成処理では、処理シーケンサ４１１は、アドレステーブル記憶部４１４中のアドレステーブルの全エントリに対して以下の処理を行う。すなわち、最新論理アドレス空間に対応するフラグに「変更有り」が記録されていた場合には、そのエントリのスナップショット論理アドレス空間に対応する物理ブロック番号を、未使用ブロックとしてアドレステーブル記憶部４１４中の未使用ブロックテーブルに登録する。また、スナップショット論理アドレス空間に対応するブロック番号に最新論理アドレス空間に対応する物理ブロック番号をコピーする。このとき、フラグの値もコピーする。その後、最新論理アドレス空間に対応するフラグを「変更無し」に初期化する。

【０２５６】

例えば、図３１（ｃ）に示すエントリでは、最新論理アドレス空間に対応するフラグに「変更有り」が記録されている。この場合、スナップショット論理アド

レス空間に対応する物理ブロック番号「ａａａ」を未使用ブロックとして未使用ブロックテーブルに登録する。そして、スナップショット論理アドレス空間に対応するブロック番号に、最新論理アドレス空間に対応する物理ブロック番号「ｂｂｂ」をコピーする。そして、最新論理アドレス空間に対応するフラグ「変更有り」も同様にコピーする。さらに、最新論理アドレス空間に対応するフラグを「変更無し」に初期化する。すると、エントリの状態は、図３１（ｃ）に示す状態から、図３１（ｄ）に示す状態になる。

【０２５７】

また、最新論理アドレス空間に対応するフラグが「変更無し」である場合、直前の再開可能ポイント以降その物理ブロック番号に対応するアドレスに格納されたデータが変更されていないことを意味する。従って、そのエントリには何も処理を行わない。

【０２５８】

最新論理アドレス空間に対応するフラグが「変更有り」となっている全エントリに対して処理を終了した後、処理シーケンサ４１１は、通信部４１２を用いてスナップショット作成コマンドに対する応答を送信する。

【０２５９】

次に、災害発生時の動作を説明する。災害等によってホスト３００が使用できなくなった場合には、ストレージ４０１を待機系から正常系にする。また、ホスト３０３が、ホスト３００から処理を引き継ぐ。

【０２６０】

障害検知からアプリケーション再開までのホスト３０３の動作を説明する。まず、ホスト３０３中のホスト監視部３１３（図２０参照）がホスト３００の異常を検出する。ホスト監視部３１３は、ホスト３００の異常を検出すると、待機系であるストレージ４０１に対してスナップショット復帰コマンドを発行する。そして、スナップショット復帰コマンドに対する応答を待ち、応答を受けたら、ホスト監視部３１３は、アプリケーション３１０ａ，３１０ｂを実行させる。以後、ホスト３０３からのデータの書き込みおよび読み出しは、ストレージ４０１に対して行われる。

【０２６１】

なお、ホスト３０３中のホスト監視部３１３は、ホスト３００と常時あるいは定期的に通信を行っている。ホスト監視部３１３が、一定時間ホスト３００と通信できなくなった場合、あるいは、ホスト３００から異常が報告された場合に、ホスト監視部３１３は、ホスト３００において災害が発生したと認識する。

【０２６２】

次に、ストレージ４０１が、スナップショット復帰コマンドを受けた場合の動作を説明する。ストレージ４０１にスナップショット復帰コマンドが到着すると、ストレージ４０１のストレージコントローラ４１０における通信部４１２にスナップショット復帰コマンドが入力される。通信部４１２は、スナップショット復帰コマンドを受け取ると、処理シーケンサ４１１にスナップショット復帰コマンドを渡し、スナップショット復帰処理の開始を指示する。

【０２６３】

スナップショット復帰処理では、処理シーケンサ４１１は、アドレステーブル記憶部中４１４のアドレステーブルの全エントリに対して以下の処理を行う。すなわち、最新論理アドレス空間に対応するフラグに「変更有り」が記録されていた場合には、その最新論理アドレス空間に対応する物理ブロック番号を未使用ブロックとしてアドレステーブル記憶部４１４の未使用ブロックテーブルに登録する。その結果、その物理ブロック番号に対応する領域は未使用状態として解放される。次いで、記憶媒体へのデータ格納状況を示す格納情報（最新論理アドレス空間に対応する物理ブロック番号およびフラグ）を、直前のスナップショット作成時の状態に戻す。すなわち、最新論理アドレス空間に対応するブロック番号にスナップショット論理アドレス空間に対応する物理ブロック番号をコピーする。また、最新論理アドレス空間に対応するフラグを「変更有り」から「変更無し」にする。

【０２６４】

例えば、スナップショット復帰処理開始時に、あるエントリが図３１（ｃ）に示す状態であったとする。この最新論理アドレス空間に対応するフラグは「変更有り」となっている。従って、処理シーケンサ４１１は、その最新論理アドレス

空間に対応する物理ブロック番号「ｂｂｂ」を未使用ブロックとして未使用ブロックテーブルに登録する。次いで、最新論理アドレス空間に対応するブロック番号「ｂｂｂ」にスナップショット論理アドレス空間に対応する物理ブロック番号をコピーする。この結果、最新論理アドレス空間に対応するブロック番号は「ａａａ」になる。また、最新論理アドレス空間に対応するフラグを「変更有り」から「変更無し」にする。この結果、格納情報（最新論理アドレス空間に対応する物理ブロック番号およびフラグ）は、直前のスナップショット作成時の状態に戻る。このように、スナップショット復帰処理を行うことにより、エントリは直前の再開可能ポイントにおける状態に復帰する。

【０２６５】

また、最新論理アドレス空間に対応するフラグが「変更無し」である場合、直前の再開可能ポイント以降その物理ブロック番号に対応するアドレスに格納されたデータが変更されていないことを意味する。従って、そのエントリには何も処理を行わない。

【０２６６】

本実施の形態では、待機系のストレージ４０１はストレージ４００から受信したデータを、新たに割り当てた物理ブロック番号に対応するアドレスに格納する。そして、ストレージ４０１は、再開可能ポイントにおいてスナップショット作成コマンドを受信すると、その新たに割り当てた物理ブロックの情報をスナップショットとして保持する。また、再開可能ポイントの前に、正常系のホストに異常が発生した場合には、新たに割り当てた物理ブロック番号を未使用状態にする。新たに割り当てた物理ブロック番号を未使用の状態にすれば、ホスト３０３は、アプリケーション３１０ａ，３１０ｂの処理を再開できる。すなわち、ホスト３００に異常が生じたとしても、ホスト３０３が処理を再開するまでの時間は短くて済む。

【０２６７】

また、正常系ストレージ４００は、再開可能ポイントと再開可能ポイントとの間で記憶媒体１０１に書き込んだデータを一度に待機系ストレージ４０１に送信するのではなく、ホスト３００からwrite コマンドを受信したタイミング毎に送

信する。従って、大量のデータを一度に送信しないので、ホスト４０１とのデータ転送時間が少なくてすむ。

【０２６８】

次に、第８の実施の形態の変形例について説明する。図３２および図３５に示された処理例では、再開可能ポイント通知処理が行われていた場合にはwrite コマンドの発行は待たされていた（ステップＳ５０５参照）。また、write 処理のうちで完了していないものがある場合にはスナップショット作成コマンドの発行が待たされていた（ステップＳ５６０参照）。これに対し、本変形例では、ストレージ４００は、ホスト３００からwrite コマンドを受信した場合、再開可能ポイント通知処理の状態によらずにストレージ４０１にwrite コマンドを送信する。すなわち、図３２に示すステップＳ５０４の後、即座にステップＳ５０６以降の処理を開始する。また、ストレージ４００は、ホスト３００から再開可能ポイント通知を受けた場合には、write 処理の状況によらず即座にストレージ４０１にスナップショット作成コマンドが発行される。すなわち、再開可能ポイント通知を受けた場合、即座に図３５に示すステップＳ５６２以降の処理を開始する。

【０２６９】

本変形例では、ストレージ４００は、ストレージ４０１に対して送信する全てのコマンドに発行ＩＤを付加する。また、待機系のストレージ４０１は、処理済発行ＩＤ情報を保持する。処理済発行ＩＤ情報は、ストレージ４００から受信した各コマンドのどのコマンドまでの処理が完了したのかを示す情報である。例えば、処理済発行ＩＤ情報の内容が「５３」であるならば、発行ＩＤ「５３」までのコマンドに対する処理が完了したことを意味する。

【０２７０】

図３６は、第８の実施の形態の変形例において、write コマンドを受信したストレージ４０１の処理シーケンサ４１１の動作を示すフローチャートである。処理シーケンサ４１１は、まず、受け取るデータに必要な領域をバッファメモリ１０６に確保する（ステップＳ５８０）。また、write コマンドで指定された書き込み先の論理アドレスとwrite 処理であることを指定して、ＬＢＡ管理部４１３に論理アドレスを物理アドレスに変換させる（ステップＳ５８１）。ステップＳ

５８１において、ＬＢＡ管理部４１３は、ステップＳ５２１の場合と同様にエントリに対して処理を行い、物理アドレスを処理シーケンサ４１１に渡す。

【０２７１】

次いで、処理シーケンサ４１１は、準備完了の通知をデータ転送元のストレージ（この例ではストレージ４００）に送るように通信部４１２に指示する（ステップＳ５８２）。通信部４１２は、指示に応じて、データ転送元のストレージに準備完了の通知を送信する。そして、データ転送元のストレージからデータが届くのを待ち（ステップＳ５８３）、データが届いて通信部４１２からデータを格納すべきバッファメモリ１０６の領域の問い合わせを受けると、ステップＳ５８０で確保した領域を通信部４１２に知らせる（ステップＳ５８４）。

【０２７２】

さらに、データ転送元のストレージからのデータのバッファメモリ１０６への格納の完了を待ち（ステップＳ５８５）、全てのデータがバッファメモリ１０６に格納されたことが通信部４１２から通知されると、処理シーケンサ４１１は、処理済発行ＩＤ情報の値が、処理対象のwrite コマンドに付加された発行ＩＤの一つ前の値になるまで待つ（ステップＳ５８６）。例えば、発行ＩＤ「５４」のwrite コマンドを受信して、ステップＳ５８０からステップＳ５８５までの動作を行ったとする。このとき、発行ＩＤ「５３」までの各コマンド（write コマンドやスナップショット作成コマンド等）を受信しているとは限らない。この場合、ステップＳ５８６では、発行ＩＤ「５３」までの各コマンドを受信し、その各コマンドに応じた処理を全て完了させるまで、発行ＩＤ「５４」に対する処理を中断する。処理済発行ＩＤ情報が「５３」になったならば、発行ＩＤ「５４」に対する処理を再開し、ステップＳ５８７に移行する。

【０２７３】

処理済発行ＩＤ情報の値が発行ＩＤの一つ前の値になったら、処理シーケンサ４１１は、ＩＯスケジューラ１０４に対して、ステップＳ５８１で変換した物理アドレスに対応した領域に、バッファメモリ１０６に格納されたデータを書き込む指示（書き込み要求）を登録する。ＩＯスケジューラ１０４は、書き込み要求に応じて、媒体制御部１０５に、書き込み対象のデータの書き込み指示を行う。

媒体制御部１０５は、登録内容に応じてバッファメモリ１０６から記憶媒体１０１へのデータの書き込み処理を行う（ステップＳ５８７）。

【０２７４】

そして、媒体制御部１０５からの書き込み完了通知を待ち（ステップＳ５８８）、媒体制御部１０５からの書き込み完了通知を受けると、データ転送元のストレージに完了通知を行う（ステップＳ５８９）。また、ステップＳ５８９では、write コマンドに付加された発行ＩＤを処理済発行ＩＤ情報に反映させる。例えば、発行ＩＤ「５４」のwrite コマンドについてステップＳ５８９までの処理を行ったならば、処理済発行ＩＤ情報の情報を「５３」から「５４」に更新する。

【０２７５】

また、第８の実施の形態の変形例では、ストレージ４０１がスナップショット作成コマンドを受信した場合には、以下のように動作する。すなわち、ストレージ４０１にスナップショット作成コマンドが到着すると、ストレージ４０１のストレージコントローラ４１０の通信部４１２にスナップショット作成コマンドが入力される。通信部４１２は、スナップショット作成コマンドを受け取ると、処理シーケンサ４１１にスナップショット作成コマンドを渡し、スナップショット作成処理の開始を指示する。

【０２７６】

スナップショット作成処理では、処理済発行ＩＤ情報後が、処理対象の要求に付加された発行ＩＤの一つ前の値になるまで待つ。すなわち、ステップＳ５８６の場合と同様に、スナップショット作成コマンドよりも一つ前に発行されたコマンドに応じた処理を全て完了させるまで、そのスナップショット作成コマンドに応じた処理を進めずに待つ。

【０２７７】

処理済発行ＩＤ情報の値が発行ＩＤの一つ前の値になったら、処理シーケンサ４１１は、アドレステーブル記憶部中４１４のアドレステーブルの全エントリに対して以下の処理を行う。すなわち、最新論理アドレス空間に対応するフラグに「変更有り」が記録されていた場合には、その最新論理アドレス空間に対応する物理ブロック番号を未使用ブロックとしてアドレステーブル記憶部４１４の未使

用ブロックテーブルに登録する。次いで、最新論理アドレス空間に対応するブロック番号にスナップショット論理アドレス空間に対応する物理ブロック番号をコピーする。また、最新論理アドレス空間に対応するフラグを「変更有り」から「変更無し」にする。

【０２７８】

また、最新論理アドレス空間に対応するフラグが「変更無し」である場合、直前の再開可能ポイント以降その物理ブロック番号に対応するアドレスに格納されたデータが変更されていないことを意味する。従って、そのエントリには何も処理を行わない。

【０２７９】

以上の処理全てが終了したならば、処理シーケンサ４１１は、処理済発行ＩＤ情報の値をスナップショット作成コマンドに付加された発行ＩＤに変更する。そして、通信部４１２を用いてスナップショット作成コマンドに対する応答をストレージ４００に送信する。

【０２８０】

本変形例では、ストレージ４０１が発行ＩＤの順番に各コマンドの処理を進めていく点が、既に説明した第８の実施の形態と異なる。しかし、エントリに対する処理自体に相違点はない。従って、ホスト３００に異常が生じたとしても、ホスト３０３が処理を再開するまでの時間は短くて済む。また、ストレージ４００，４０１間でのデータ転送時間も短くて済む。

【０２８１】

第８の実施の形態において、書き込み要求手段およびスナップショット作成要求手段は、ストレージ４００の処理シーケンサ４１１および通信部４１２によって実現される。スナップショット作成手段は、ストレージ４０１の処理シーケンサ４１１、ＬＢＡ管理部４１３およびアドレステーブル記憶部４１４によって実現される。再開可能ポイント通知手段は、ホスト３００の再開可能ポイント通知部３１２によって実現される。

【０２８２】

実施の形態９．

図３７は、本発明によるデータ複製システムの第９の実施の形態を示すブロック図である。図３７に示すデータ複製システムにおいて、正常系のストレージ４００が、ストレージ４００を使用する正常系のホスト６００とローカルに接続されている。また、待機系のストレージ４０１が、ストレージ４０１を使用する待機系のホスト６０１とローカルに接続されている。ストレージ４００は、ネットワーク１３を介してストレージ４０１に接続されている。また、ホスト６００は、ネットワーク６０２を介してホスト６０１に接続されている。

【０２８３】

なお、ストレージ４００，４０１の構成および動作は、第８の実施の形態におけるスナップショット機能を有するストレージ４００，４０１の構成および動作と同じである（図２９参照）。また、ストレージ４００，４０１に代えて、第７の実施の形態におけるストレージ３０１，３０２（図２１参照）を用いてもよい。

【０２８４】

また、ネットワーク１３，６０２は、１つのネットワークであってもよい。ただし、ホスト６００，６０１が接続されるネットワーク６０２は、専用回線とすることが好ましい。

【０２８５】

図３８は、ホスト６００の構成例を示すブロック図である。図３８に示すように、ホスト６００において、単数あるいは複数のアプリケーションが動作する。ここでは、２つのアプリケーション６０３ａ，６０３ｂを例示する。アプリケーション６０３ａ，６０３ｂは、ＩＯ管理部３１１を用いて、ストレージ４００中のデータにアクセスを行う。また、ＩＯ管理部３１１は、ストレージ４００に再開可能ポイントを通知するための再開可能ポイント通知部３１２を有する。また、ホスト６０１にアプリケーション６０３ａ，６０３ｂの実行イメージを転送し、再開可能ポイント通知部３１２に再開可能ポイントを通知する実行イメージ転送部６０４が備えられている。

【０２８６】

本実施の形態におけるアプリケーション６０３ａ，６０３ｂは、再開機能を有

さないアプリケーションである。すなわち、ストレージの記憶媒体１０１のデータ記録状態が所定の状態になっていれば処理を再開できるような機能を有していないアプリケーションである。

【０２８７】

図３９は、ホスト６０１の構成例を示すブロック図である。図３９に示すように、ホスト６０１には、実行イメージを保存する実行イメージ保存部６０６と、ホスト６００から送られてきた実行イメージを受け取り実行イメージ保存部６０６に保存する実行イメージ受信部６０５と、ホスト６００の状況を監視するホスト監視部６０８と、実行イメージ保存部６０６中の実行イメージを元にアプリケーションを再開させるアプリケーション再開部６０７とが備えられている。実行イメージ保存部６０６は、揮発性半導体メモリ、不揮発性半導体メモリ、磁気ディスク、光磁気ディスク、光ディスク等のデータを保存する媒体である。

【０２８８】

実行イメージは、稼働系の処理実行状態を示す情報であり、各アプリケーションによって実行されるプロセスを他のホストで実行させるのに必要な情報である。実行イメージには、例えば、各プロセスの仮想アドレス空間中のデータ、プロセス管理情報であるレジスタの値、プログラムカウンタの値およびプロセスの状態等である。プロセスが使用しているファイルやプロセス間通信を復元するための情報等が含まれる。また、並列処理によって一つのプロセスを複数のスレッドで実現する場合には、各スレッドのレジスタの値、プログラムカウンタの値、プログラム状態ワードおよび各種フラグ等も実行イメージに含まれる。さらに、実行イメージは、対象プロセスが使用するカーネル中の通信バッファの内容等を含む場合もある。なお、スレッドとは、並列処理における処理単位である。

【０２８９】

次に、動作について説明する。まず、実行イメージ転送時の動作を説明する。正常系のホスト６００の実行イメージ転送部６０４には、例えば、ユーザから実行イメージ転送指示が入力される。この指示が入力されるタイミングは、任意のタイミングでよい。指示が入力されると、実行イメージ転送部６０４は、アプリケーション６０３ａ，６０３ｂの実行イメージを取得する。次に、実行イメージ

転送部６０４は、ＩＯ管理部３１１の再開可能ポイント通知部３１２に、ストレージ４００に対して再開可能ポイントを通知するように指示する。そして、再開可能ポイント通知部３１２は、ストレージ４００に再開可能ポイントを通知する。次いで、実行イメージ転送部６０４は、ホスト６０１の実行イメージ受信部６０５に対して、取得した実行イメージを転送する。ホスト６０１において、実行イメージ受信部６０５は、ホスト６００の実行イメージ受信部６０６から実行イメージを受け取ると、受け取った実行イメージを実行イメージ保存部６０６に保存する。

【０２９０】

ストレージ４００は、実行イメージの転送が指示された時点において再開可能ポイント通知を受信し、その通知に応じて動作する。従って、待機系のストレージ４０１は、正常系のホストにおいて実行イメージ転送が指示された時点におけるスナップショットを作成する。

【０２９１】

ストレージとして第７の実施の形態と同様のストレージ３０１，３０２（図１９参照）を用いる場合、待機系のストレージ４０１は、記憶媒体の状態を、正常系のホストにおいて実行イメージ転送が指示された時点における状態に維持する。

【０２９２】

なお、実行イメージの容量は、数百ＭＢ以上になる場合もあり、ホスト６０１への実行イメージの転送には時間がかかる。そして、実行イメージの転送中にアプリケーション６０３ａ，６０３ｂが処理を進めるとメモリやレジスタの値が変化してしまい、転送途中で実行イメージの内容が変わってしまう。そのため、実行イメージ転送中は、アプリケーション６０３ａ，６０３ｂは処理を停止する。

【０２９３】

あるいは、ホスト６００が備えるメモリや磁気ディスク等（図３７において図示せず）に、実行イメージの情報を保存し、保存した実行イメージをホスト６０１に送信してもよい。一旦、保存完了後に、処理を進めても、保存された実行イメージの内容は保たれる。従って、この場合、ホスト６００のアプリケーション

６０３ａ，６０３ｂは処理を進めることができる。実行イメージをメモリ等に保存する時間は、実行イメージの転送時間に比べ短い。従って、実行イメージをメモリ等に保存してから転送するようにすれば、処理を停止する時間は短くて済む。

【０２９４】

ここでは、実行イメージ転送部６０４がユーザからの指示に応じて実行イメージを転送する場合を示した。実行イメージ転送部６０４が実行イメージ転送処理を開始するタイミングは、ユーザから指示が入力された時点に限定されない。例えば、一定時間間隔毎に実行イメージ転送処理を開始するようにしてもよい。また、実行イメージの転送が終了したときに再度実行イメージの転送処理を開始してもよい。あるいは、アプリケーションがアイドルになった時点、アプリケーションのアイドルタイムが一定値を越えた時点等に転送処理を開始してもよい。

【０２９５】

次に、災害発生時の動作を説明する。災害等によってホスト６００が使用できなくなった場合には、ストレージ４０１を待機系から正常系にする。また、ホスト６０１が、ホスト６００から処理を引き継ぐ。

【０２９６】

障害検知からアプリケーション再開までのホスト６０１の動作を説明する。まず、ホスト６０１中のホスト監視部６０８がホスト６００の異常を検出する。例えば、ホスト監視部６０８は、ホスト６００と常時あるいは定期的に通信を行い、一定時間ホスト６００と通信できなくなった場合に異常が生じたと認識する。あるいは、ホスト６００から異常が報告された場合に、ホスト６００において災害が発生したと認識してもよい。

【０２９７】

ホスト監視部６０８は、ホスト６００の異常を検出すると、待機系であるストレージ４０１に対してスナップショット復帰コマンドを発行する。そして、スナップショット復帰コマンドに対する応答を待ち、応答を受けたら、ホスト監視部６０８は、アプリケーション再開部６０７に、アプリケーションの再開を指示する。アプリケーション再開部６０７は、実行イメージ保存部６０６に保存されて

いる実行イメージを用いてアプリケーションアプリケーションの動作を再開させる。本例では、ストレージとして第五の実施の形態と同様のストレージ４００，４０１を用いる場合を示した。ストレージとして第７の実施の形態と同様のストレージ３０１，３０２（図１９参照）を用いる場合、ホスト監視部６０８は、スナップショット復帰コマンドの代わりに遅延データ破棄コマンドを出力すればよい。

【０２９８】

なお、アプリケーション再開部６０７は、アプリケーションの処理を再開するときに、実行イメージのうち動作するホストによって変更を必要とする情報がある場合には、復元時にその情報を変更する。

【０２９９】

第９の実施の形態では、正常系のホストにおいて実行イメージの転送が指示された時点で、正常系のストレージは再開可能ポイント通知を受信する。従って、待機系のストレージは、正常系のホストにおいて実行イメージの転送が指示された時点におけるスナップショットを作成する。あるいは、正常系のホストにおいて実行イメージの転送が指示された時点における記憶媒体の状態を維持する。さらに、待機系のホストは、実行イメージの転送が指示された時点における正常系ホストの実行イメージを保持する。従って、アプリケーションによって再開機能が実現されていなくても、待機系のストレージは、所定の時点における正常系のストレージの状態を再現でき、待機系のホストは、その時点における実行イメージを再現できるので、待機系で処理を迅速に再開することができる。

【０３００】

第９の実施の形態において、実行イメージ転送部６０４は、実行イメージを全て転送するのではなく、前回転送した実行イメージとの差分のみを転送してもよい。このように、前回転送した実行イメージから変更された部分のみを転送するようにすれば、転送時間を短縮化することができる。

【０３０１】

仮想記憶をコンピュータシステムでは、使用される確率の高い論理アドレスと物理アドレスとの変換テーブルを用いることが多い。この変換テーブルは、ＴＬ

Ｂ（Translation Look-aside Buffer ）と呼ばれる。ＴＬＢでは、変換される各アドレス毎に、そのアドレスのデータが書き換えられたか否かを示す更新情報を管理する。この更新情報は、一般に、ダーティ（dirty）フラグまたはダーティビットと呼ばれている。例えば、「コンピュータ・アーキテクチャ　設計・実現・評価の定量的アプローチ（デイビット・パターソン，ジョン・ヘネシー著、富田眞治　他２名訳、日経ＢＰ社，１９９４年２月，ｐ．４４２－４４３）」では、この更新情報はダーティビットとして記載されている。実行イメージ転送部６０４は、変更があったことを示すダーティフラグに対応する情報を実行イメージとして送信すればよい。

【０３０２】

また、仮想記憶において、プログラムの実行に必要なページを実記憶装置に配置することをページインとよび、ページを実記憶装置から外して仮想記憶装置に配置することをページアウトとよぶ。実行イメージ転送部６０４は、実行イメージの転送を指示された場合には、ページインしているメモリのデータをホスト６０１に送信し、その後、ページアウトが発生したタイミングで、ページアウトしたメモリのデータをホスト６０１に送信してもよい。ページアウトしたメモリのデータは、次にページインするまで変更されることがない。そのため、このように実行イメージを送信することで、実行イメージの送信量を減らすことができる。

【０３０３】

第９の実施の形態において、実行イメージ転送手段は、実行イメージ転送部６０４によって実現される。実行イメージ保存手段は、実行イメージ保存部６０６によって実現される。任意時点再開可能ポイント通知手段は、ホスト６００の再開可能ポイント通知部３１２によって実現される。

【０３０４】

図４０，４１は、クライアントサーバシステムに、第７の実施の形態から第９の実施の形態を適用した場合の構成例を示すブロック図である。通常、ホスト３００が各クライアント５００－１～５００－ｎのサーバとして機能する。災害等により、ホスト３００またはストレージ４００に以上が発生した場合、ホスト３

０３が稼働する。ホスト３０３が稼働を開始した後、各クライアントはホスト３０３をサーバとして、ホスト３０３に処理を要求する。

【０３０５】

なお、図４０に示すように、ホスト３００とホスト３０３とが直接接続されている場合には、第７の実施の形態から第９の実施の形態で説明したように、ホスト３０３がホスト３００の状態を監視する。図４１に示すように、ホスト３００とホスト３０３とが直接接続されていない場合には、各クライアント５００－１～５００－ｎがホスト３００の状態を監視し、ホスト３００に以上が発生したときに、ホスト３０３に以上の発生を通知してもよい。

【０３０６】

第７の実施の形態から第９の実施の形態において、第１の実施の形態から第４の実施の形態と同様に、中継装置を介して待機系のストレージにデータを送信するようにしてもよい。その場合、第１の実施の形態から第４の実施の形態の効果も得ることができる。また、ストレージ同士で、あるいはストレージと中継装置との間でデータを送受信するときには、第５の実施の形態または第６の実施の形態に示したように冗長化されたデータ群を送受信するようにしてもよい。その場合、第５の実施の形態または第６の実施の形態と同様の効果も得られる。

【０３０７】

【発明の効果】

本発明によれば、第一のストレージから第二のストレージに転送されるデータを中継する中継装置であって、第一のストレージが災害によって稼働できない状態になっても、稼働を継続できるとあらかじめ算定された位置に設置された中継装置を備え、第一のストレージは、データ転送の制御を行うデータ転送処理手段を含み、データ転送処理手段は、中継装置に対してデータ転送を完了したときに、第二のストレージに対してデータ転送を完了したとみなす。従って、稼働系のストレージが被災しても中継装置は被災せず、また、稼働系の上位装置は、待機系のストレージにデータが格納されるのを待たずに、次の処理を開始することができるので、データ転送に伴う処理の遅れが改善される。

【０３０８】

また、本発明によれば、送信される元データから少なくとも１つのエラー訂正のための冗長データを作成し、元データと冗長データとを別々のデータ送信単位で送信する。従って、待機系では、元データと冗長データとの集合のうちの一部から元データを復元することができ、送信過程で一部のデータが廃棄されても再度送信する必要がない。その結果、データの転送を迅速に完了させることができる。

【０３０９】

また、本発明によれば、稼働系のストレージは、データの書き込み要求が発生すると、書き込み対象のデータと遅延書き込み要求とを待機系のストレージに送信する遅延書き込み要求手段と、そのデータの状態であればアプリケーションがそのまま動作を再開可能な再開可能ポイントであることを知らせるための再開可能ポイント通知を上位装置から受けると遅延書き込み実行要求を待機系のストレージに送信する書き込み実行要求手段とを含み、待機系のストレージは、データを一時的に記憶する一時記憶手段と、受信したデータを遅延書き込み要求に応じて一時記憶手段に格納するとともに、遅延書き込み実行要求を受信すると一時記憶手段に格納されているデータを記憶媒体に格納する格納処理手段とを含む。従って、待機系の記憶媒体の状態を再開可能ポイント通知時の状態にすることができ、正常系で異常が発生したときに待機系ですぐに処理を再開することができる。

【０３１０】

また、本発明によれば、稼働系のストレージは、データの書き込み要求が発生すると、書き込み対象のデータと書き込み要求とを待機系のストレージに送信する書き込み要求手段と、そのデータの状態であればアプリケーションがそのまま動作を再開可能な再開可能ポイントであることを知らせるための再開可能ポイント通知を上位装置から受けるとスナップショット作成要求を待機系のストレージに送信するスナップショット作成要求手段とを含み、待機系のストレージは、書き込み要求を受信したら書き込み要求に対応するデータを書き込むべき領域を割り当てて記憶媒体にデータを格納し、スナップショット作成要求を受信したらスナップショットを作成するスナップショット作成手段を含む。従って、待機系の

記憶媒体の状態を再開可能ポイント通知時の状態にすることができ、正常系で異常が発生したときに待機系ですぐに処理を再開することができる。

【図面の簡単な説明】

【図１】　データ複製システムの第１の実施の形態を示すブロック図である。

【図２】　ストレージの構成例を示すブロック図である。

【図３】　中継装置の構成例を示すブロック図である。

【図４】　処理シーケンサの動作を示すフローチャートである。

【図５】　中継処理部の動作を示すフローチャートである。

【図６】　第２の実施の形態の中継処理部の動作を示すフローチャートである。

【図７】　第３の実施の形態の処理シーケンサの動作を示すフローチャートである。

【図８】　データ複製システムの第４の実施の形態を示すブロック図である。

【図９】　処理シーケンサの動作を示すフローチャートである。

【図１０】　ストレージの動作を示すフローチャートである。

【図１１】　ステップＳ２２１の処理を具体的に示すフローチャートである。

【図１２】　データ転送の例を示すタイミング図である。

【図１３】　データ複製システムの第５の実施の形態を示すブロック図である。

【図１４】　ストレージの構成例を示すブロック図である。

【図１５】　処理シーケンサの動作を示すフローチャートである。

【図１６】　通信部の動作を示すフローチャートである。

【図１７】　データ複製システムの他の構成例を示すブロック図である。

【図１８】　第６の実施の形態の処理シーケンサの動作を示すフローチャートである。

【図１９】　データ複製システムの第の実施の形態を示すブロック図である。

【図２０】　ホストの構成例を示すブロック図である。

【図２１】　ストレージの構成例を示すブロック図である。

【図２２】　同期ＩＤおよび発行ＩＤの例を示す説明図である。

【図２３】　処理シーケンサの動作を示すフローチャートである。

【図２４】　処理シーケンサが遅延write コマンドを受信したときの動作を示すフローチャートである。

【図２５】　処理シーケンサの動作を示すフローチャートである。

【図２６】　遅延データ反映処理の動作を示すフローチャートである。

【図２７】　遅延データ破棄処理を示すフローチャートである。

【図２８】　データ複製システムの第８の実施の形態を示すブロック図である。

【図２９】　ストレージの構成例を示すブロック図である。

【図３０】　アドレステーブルの一例を示す説明図である。

【図３１】　エントリの説明図である。

【図３２】　処理シーケンサの動作を示すフローチャートである。

【図３３】　処理シーケンサの動作を示すフローチャートである。

【図３４】　ＬＢＡ管理部の動作を示すフローチャートである。

【図３５】　処理シーケンサの動作を示すフローチャートである。

【図３６】　処理シーケンサの動作を示すフローチャートである。

【図３７】　データ複製システムの第９の実施の形態を示すブロック図である。

【図３８】　ホストの構成例を示すブロック図である。

【図３９】　ホストの構成例を示すブロック図である。

【図４０】　クライアントサーバシステムに本発明を適用した場合の構成例を示すブロック図である。

【図４１】　クライアントサーバシステムに本発明を適用した場合の構成例を示すブロック図である。

【符号の説明】

１０　ホスト

１１，１２　ストレージ

１３，１４　ネットワーク

１５　中継装置

１０１　記憶媒体

１０２　通信部

１０３　処理シーケンサ

１０４　ＩＯスケジューラ

１０５　媒体制御部

１０６　バッファメモリ

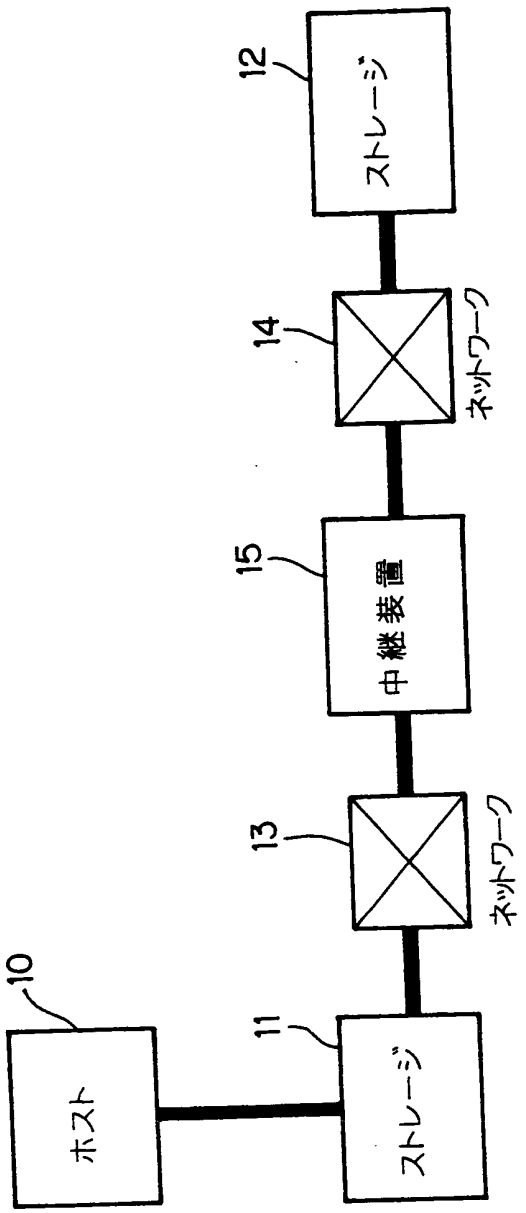
１５１　中継処理部

１５０　通信部

１５２　バッファメモリ

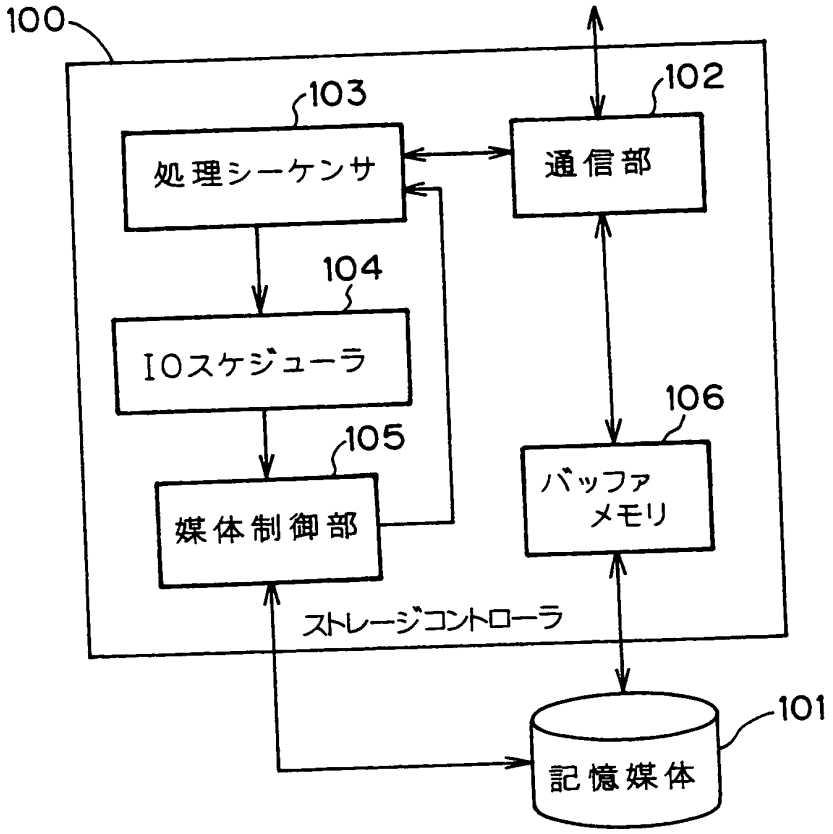
【書類名】　図面

【図１】

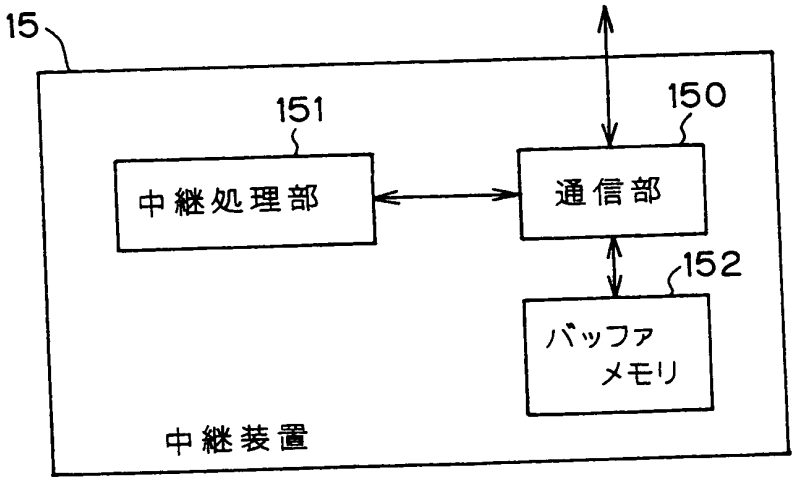


出証特２００３－３０３７４２５

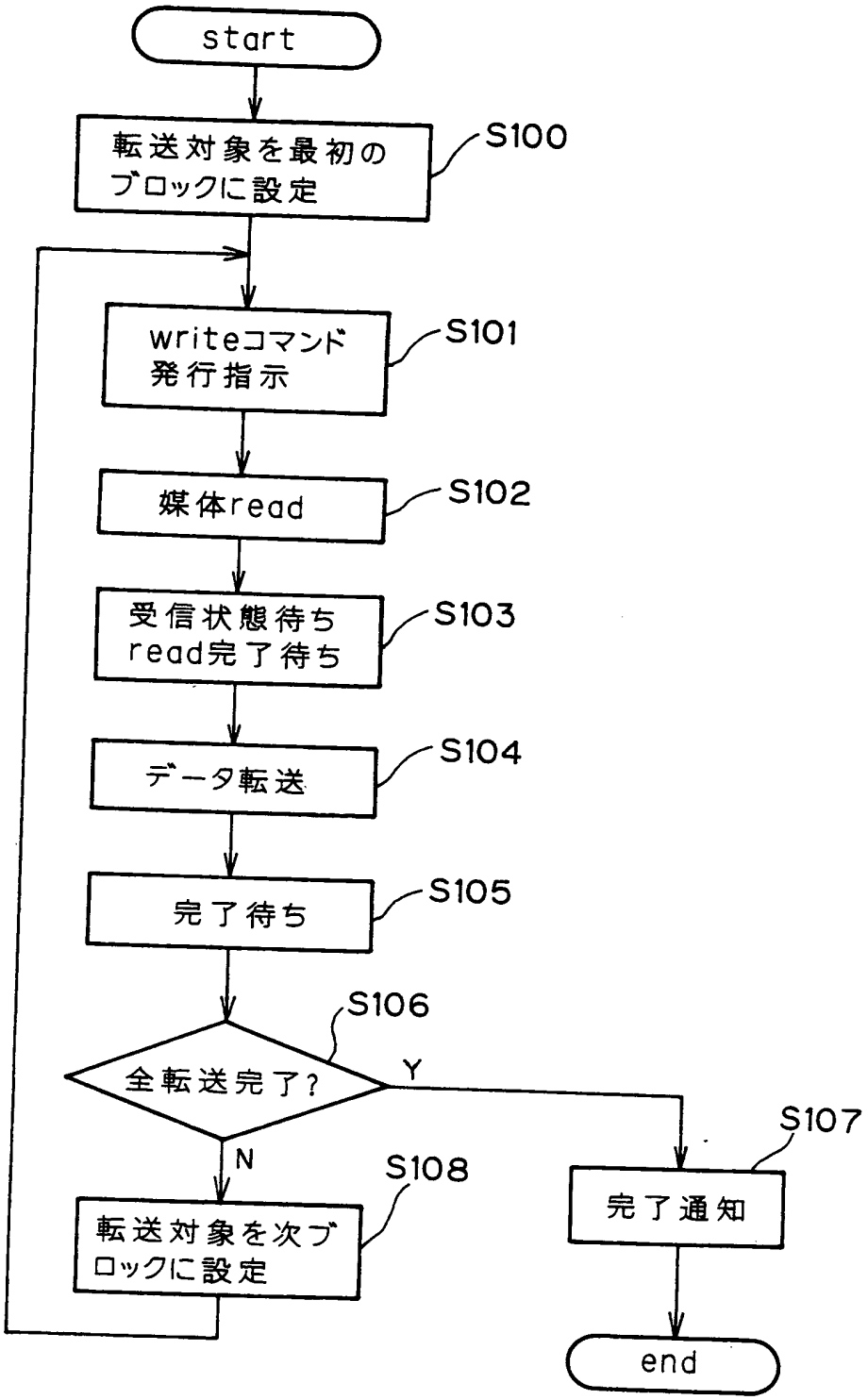
【図２】


100
103
処理シーケンサ
102
通信部
104
IOスケジューラ
106
バッファ
メモリ
105
媒体制御部
ストレージコントローラ
101
記憶媒体

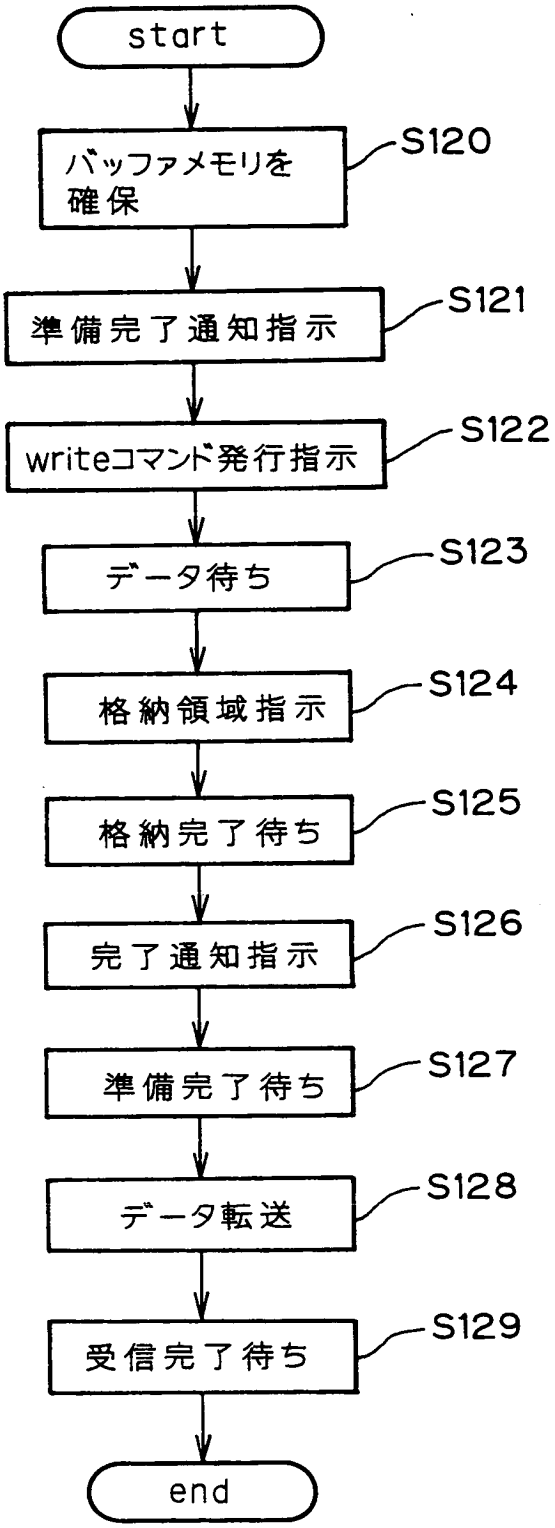
【図３】


15
151
中継処理部
150
通信部
152
バッファ
メモリ
中継装置

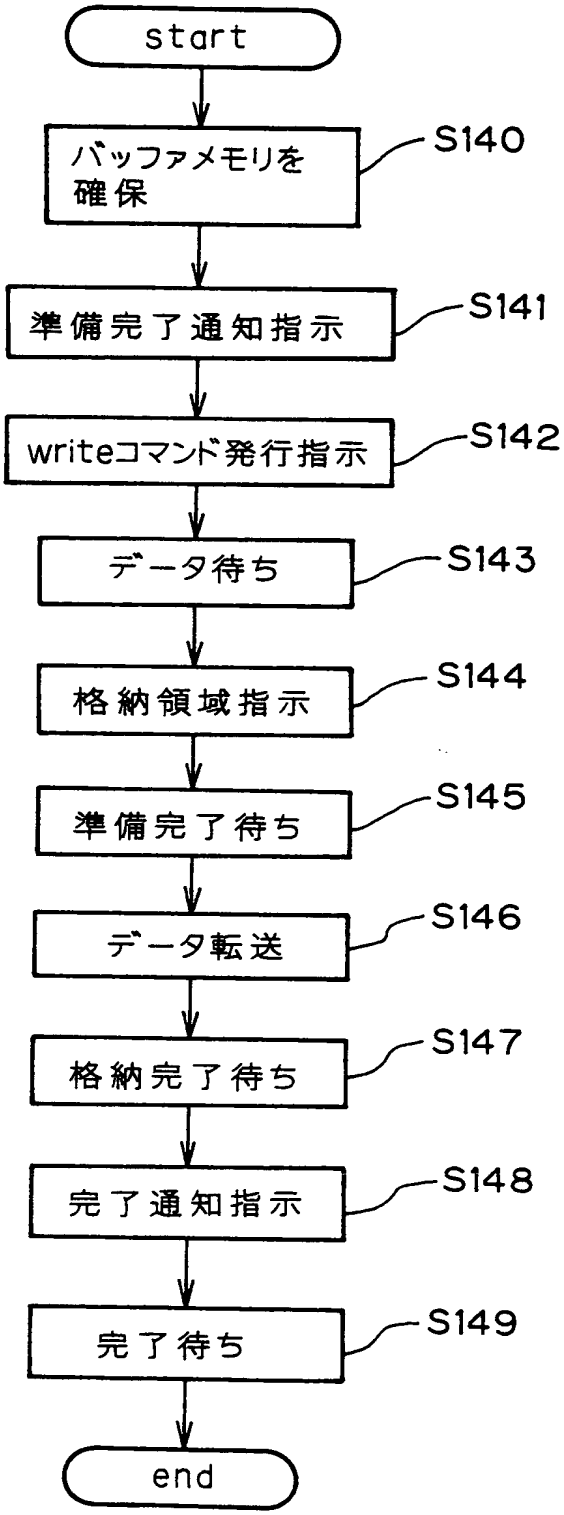
【図４】


start
転送対象を最初のブロックに設定
S100
writeコマンド発行指示
S101
媒体read
S102
受信状態待ち read完了待ち
S103
データ転送
S104
完了待ち
S105
全転送完了?
S106
Y
N
転送対象を次ブロックに設定
S108
完了通知
S107
end

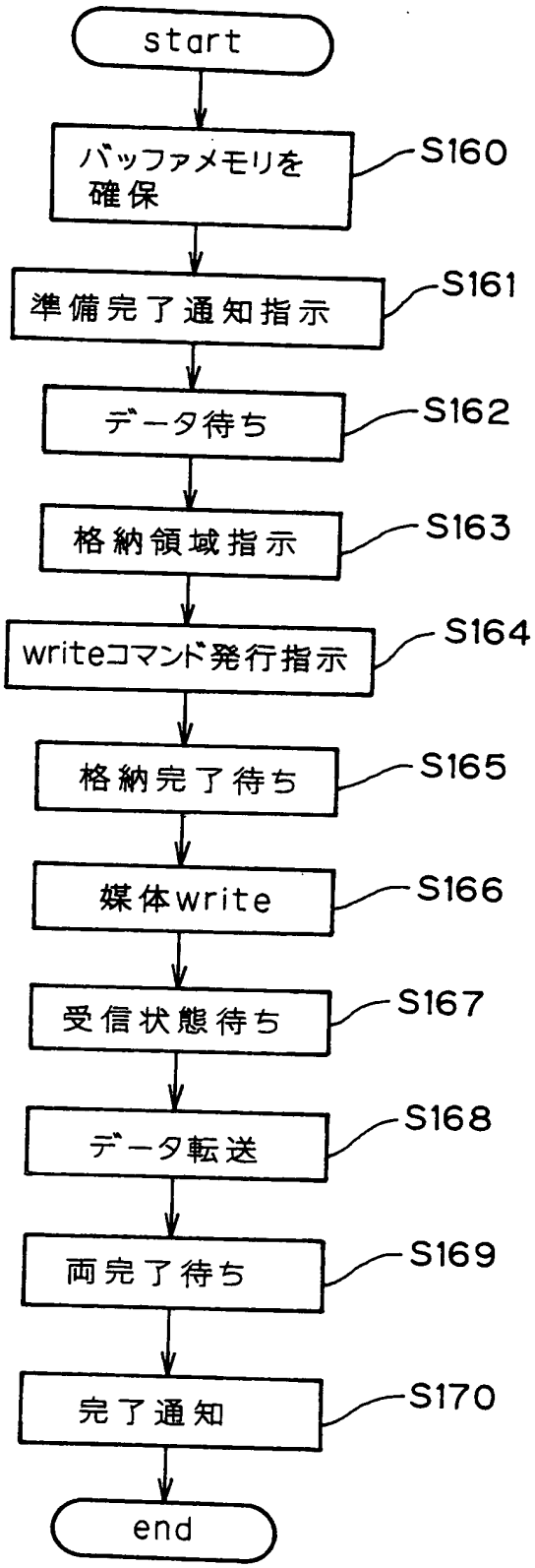
【図５】


start
バッファメモリを確保
S120
準備完了通知指示
S121
writeコマンド発行指示
S122
データ待ち
S123
格納領域指示
S124
格納完了待ち
S125
完了通知指示
S126
準備完了待ち
S127
データ転送
S128
受信完了待ち
S129
end

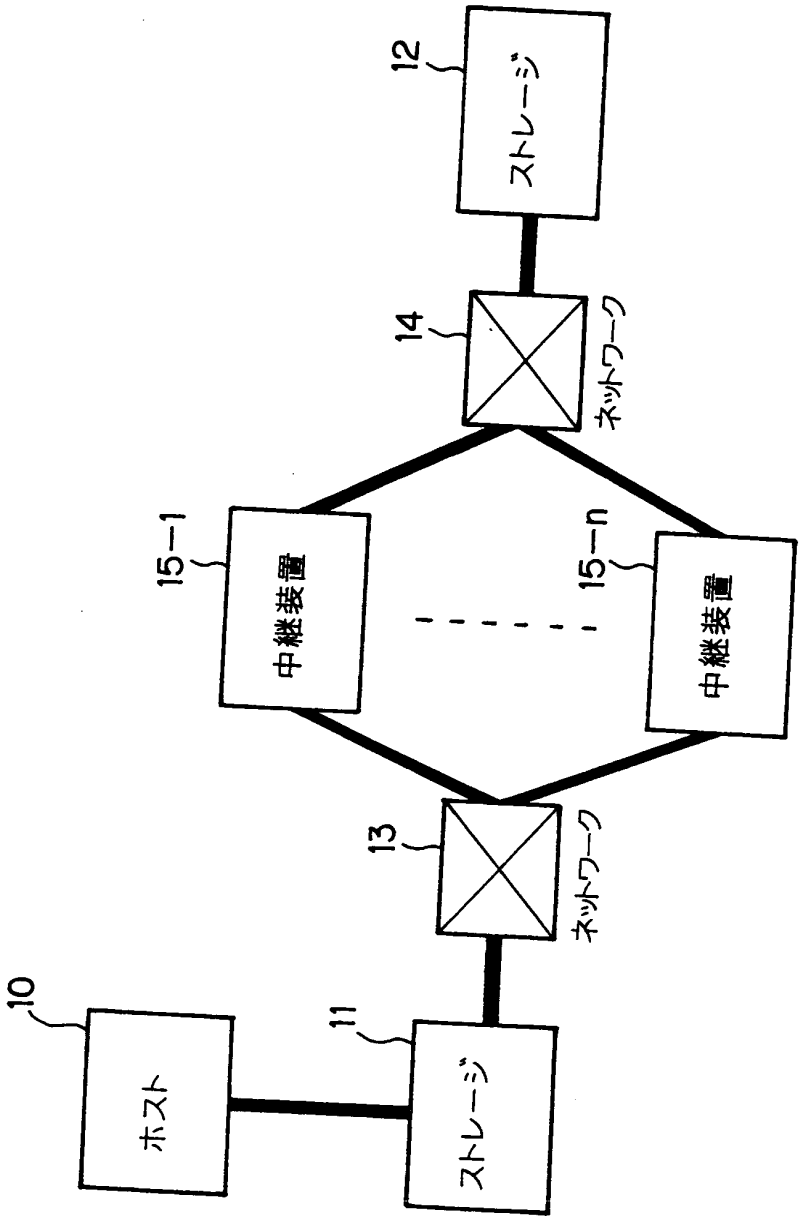
【図６】


start
バッファメモリを確保
S140
準備完了通知指示
S141
writeコマンド発行指示
S142
データ待ち
S143
格納領域指示
S144
準備完了待ち
S145
データ転送
S146
格納完了待ち
S147
完了通知指示
S148
完了待ち
S149
end

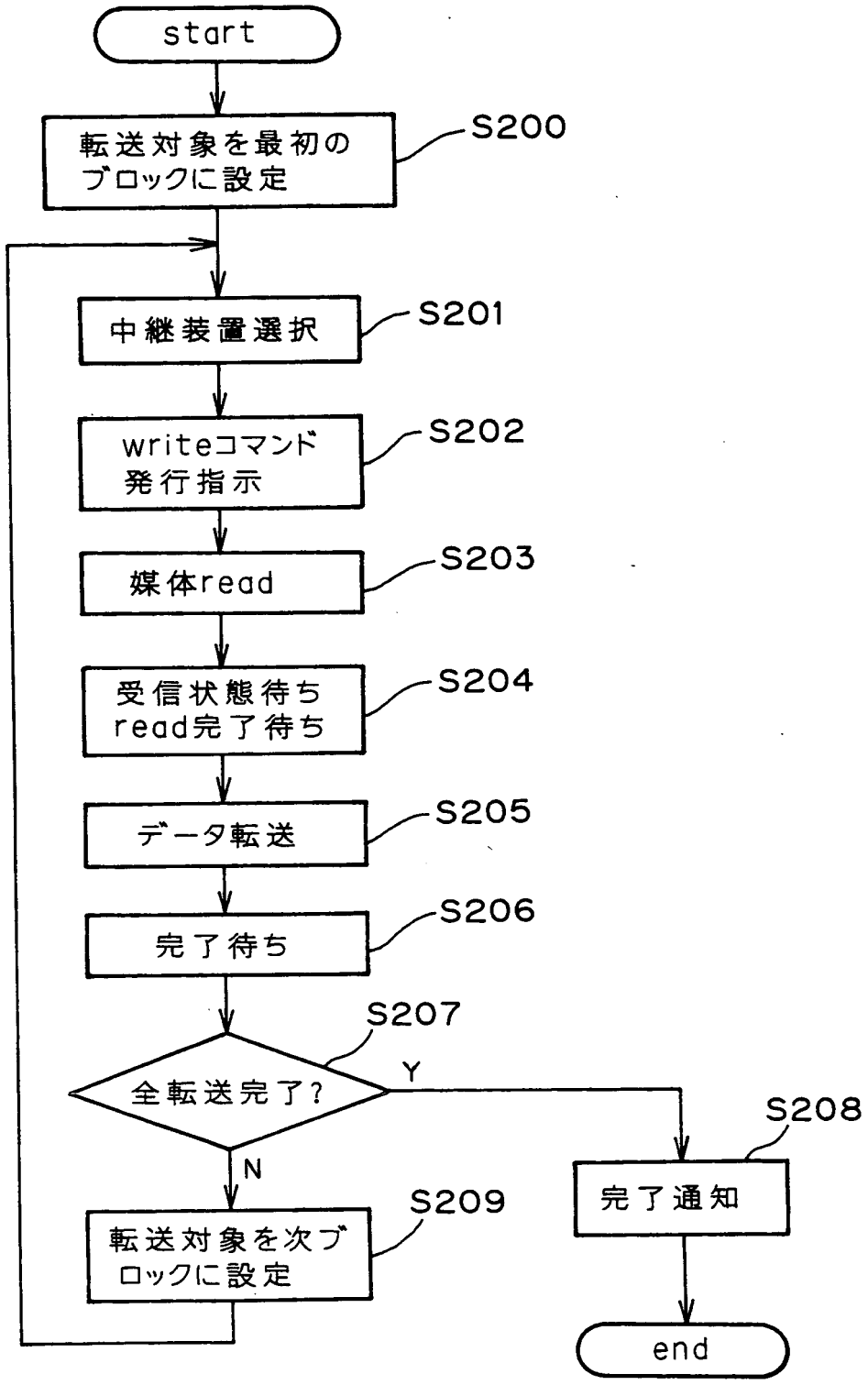
【図７】


start
バッファメモリを確保
S160
準備完了通知指示
S161
データ待ち
S162
格納領域指示
S163
writeコマンド発行指示
S164
格納完了待ち
S165
媒体write
S166
受信状態待ち
S167
データ転送
S168
両完了待ち
S169
完了通知
S170
end

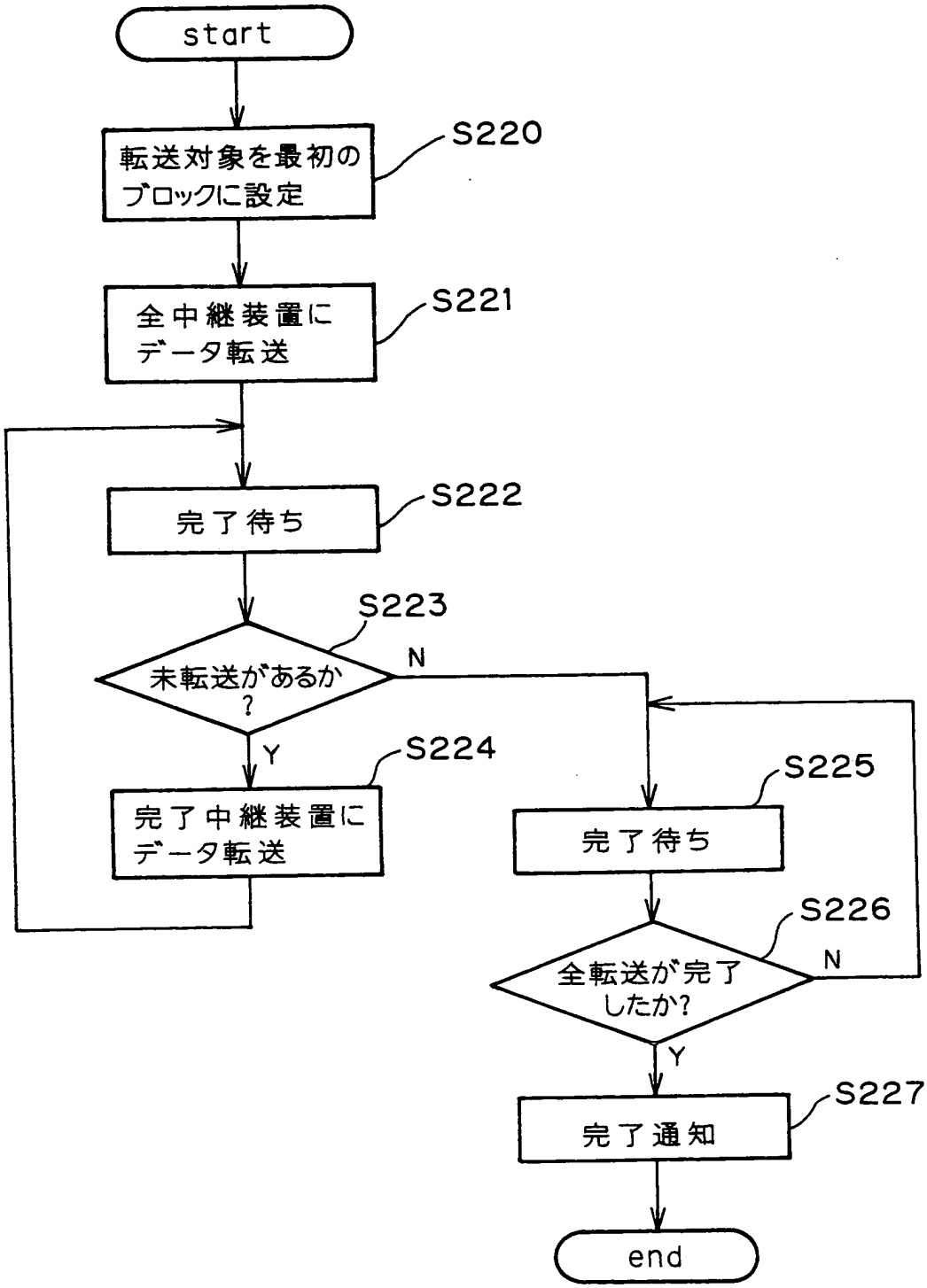
【図８】



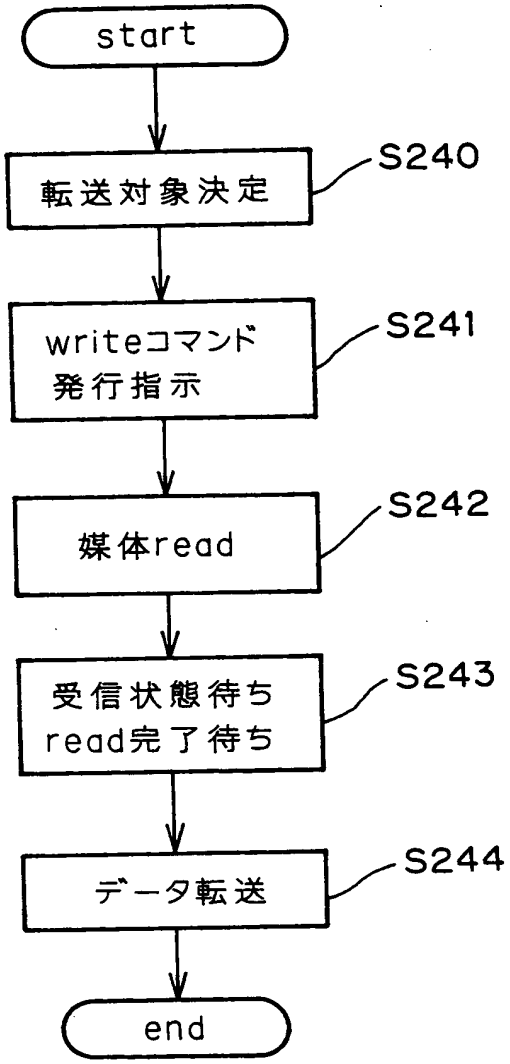
【図９】


start
転送対象を最初のブロックに設定
S200
中継装置選択
S201
writeコマンド発行指示
S202
媒体read
S203
受信状態待ち read完了待ち
S204
データ転送
S205
完了待ち
S206
S207
全転送完了?
Y
N
S208
完了通知
S209
転送対象を次ブロックに設定
end

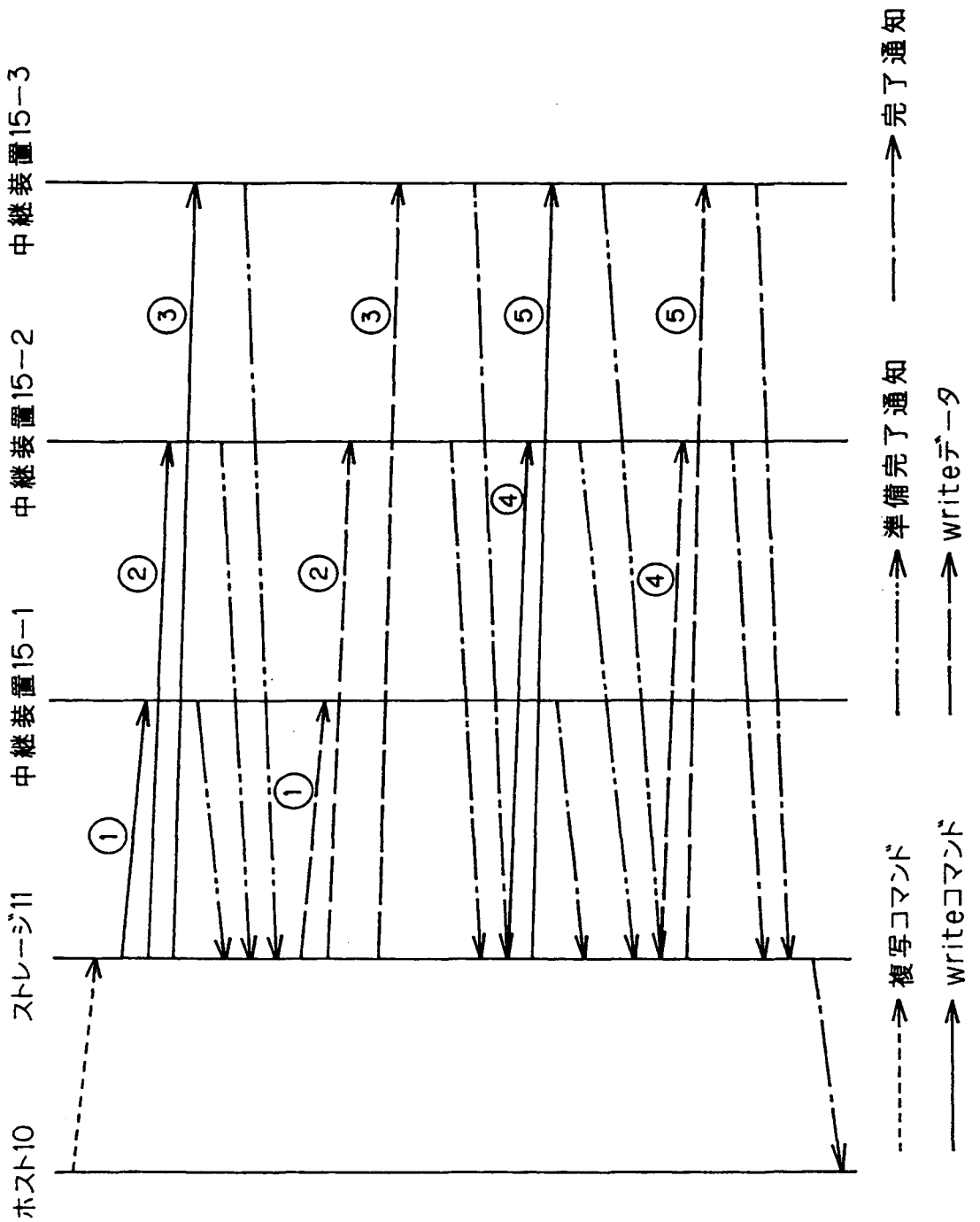
【図１０】


start
転送対象を最初のブロックに設定
S220
全中継装置にデータ転送
S221
完了待ち
S222
未転送があるか？
S223
N
Y
完了中継装置にデータ転送
S224
完了待ち
S225
全転送が完了したか？
S226
N
Y
完了通知
S227
end

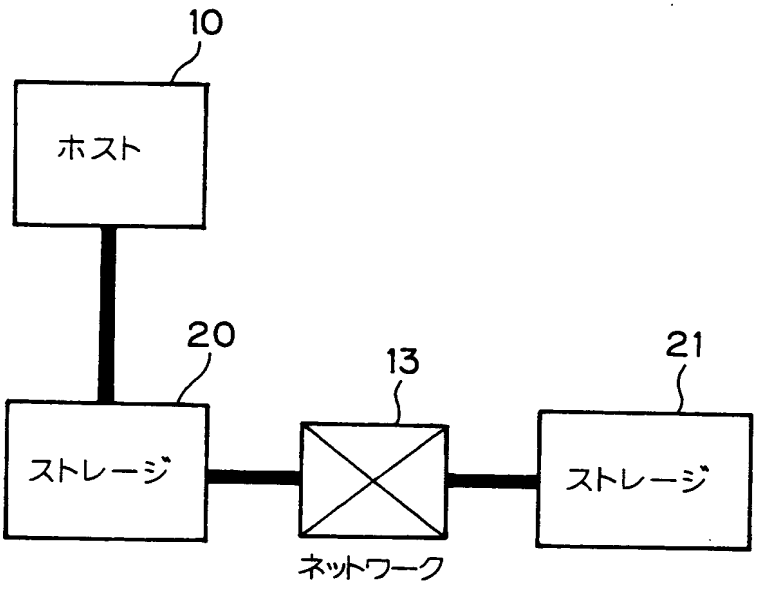
【図１１】


start
転送対象決定
S240
writeコマンド
発行指示
S241
媒体read
S242
受信状態待ち
read完了待ち
S243
データ転送
S244
end

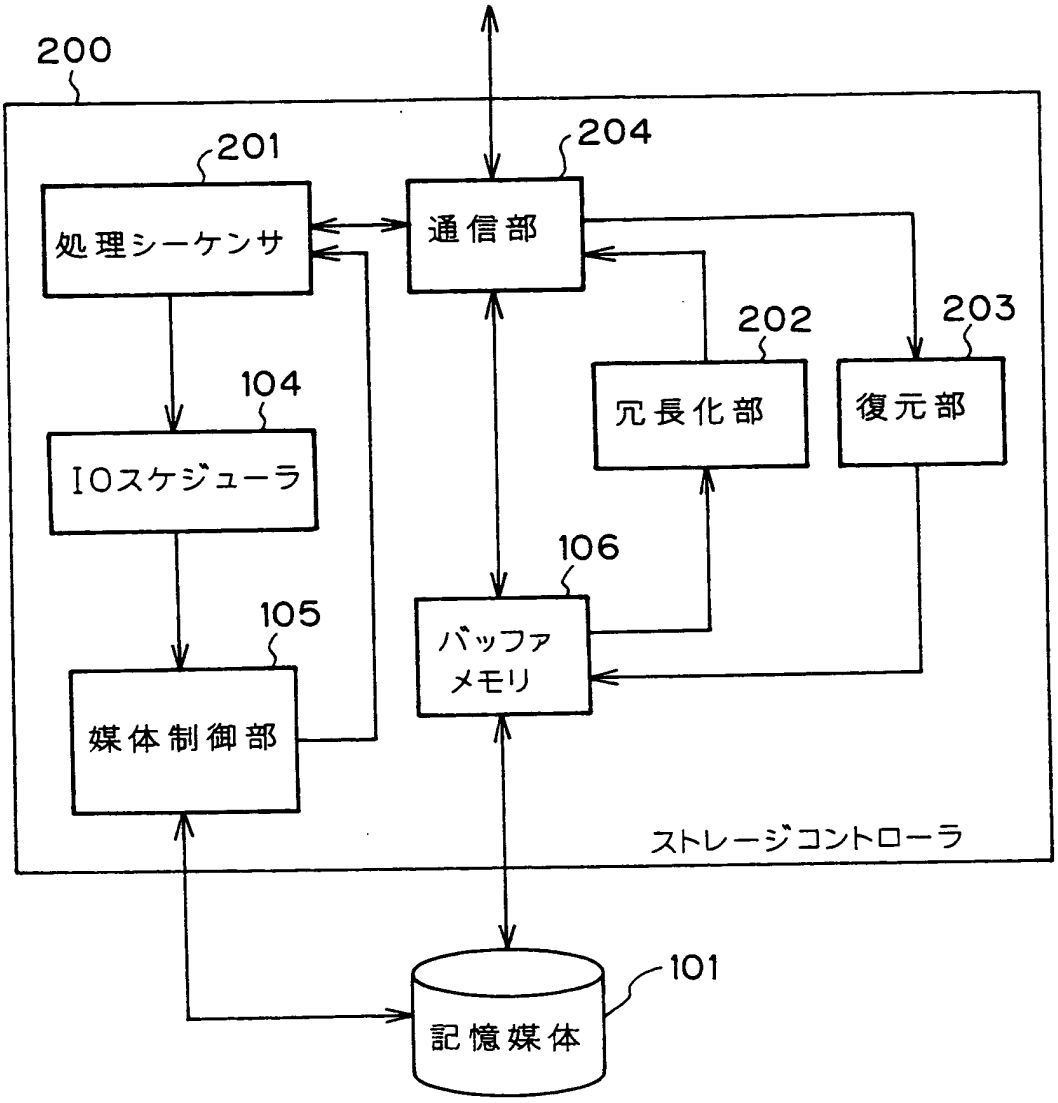
【図１２】


ホスト10
ストレージ11
中継装置15－1
中継装置15－2
中継装置15－3
①
②
③
④
⑤
複写コマンド
writeコマンド
準備完了通知
writeデータ
完了通知

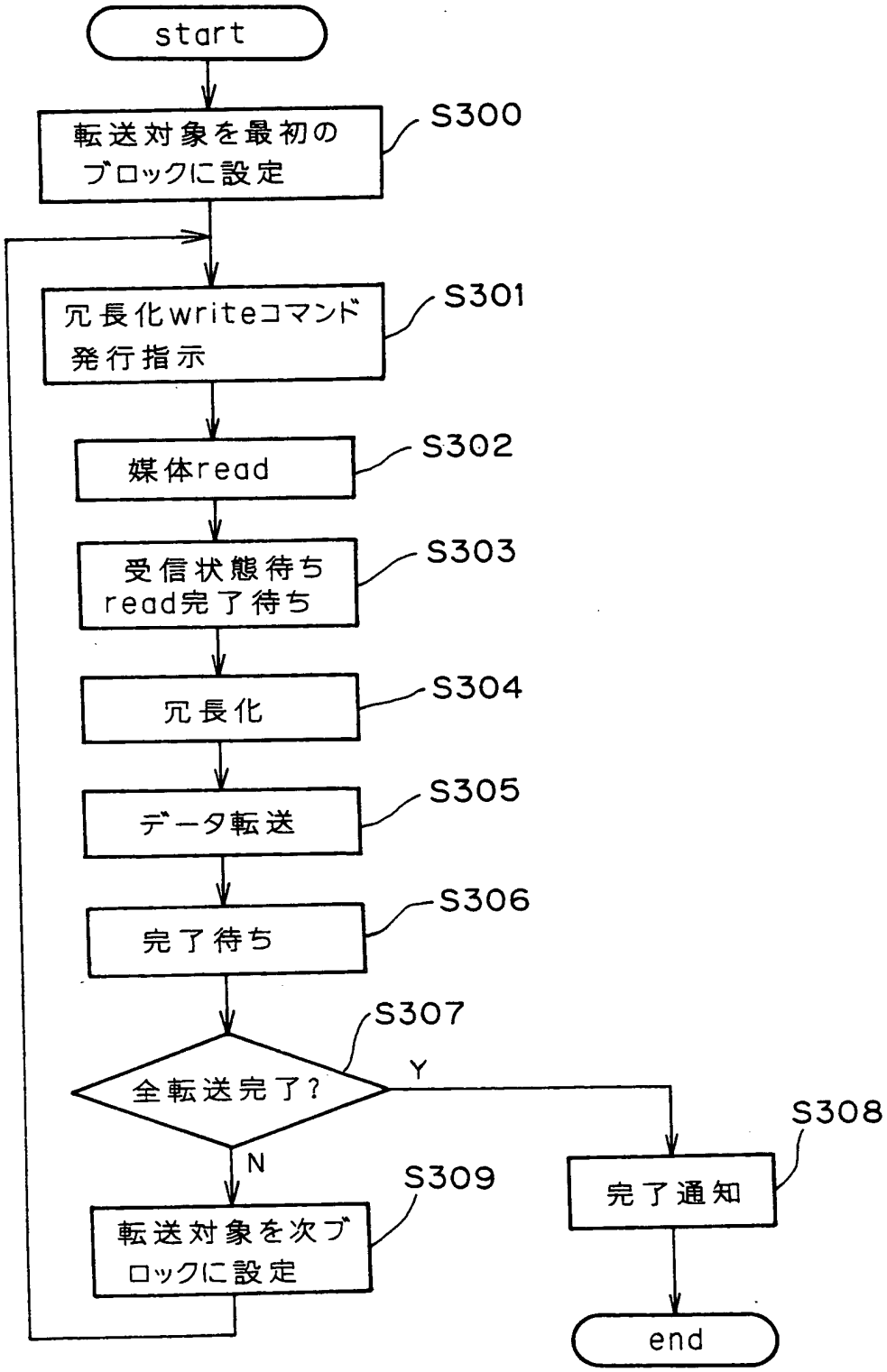
【図１３】



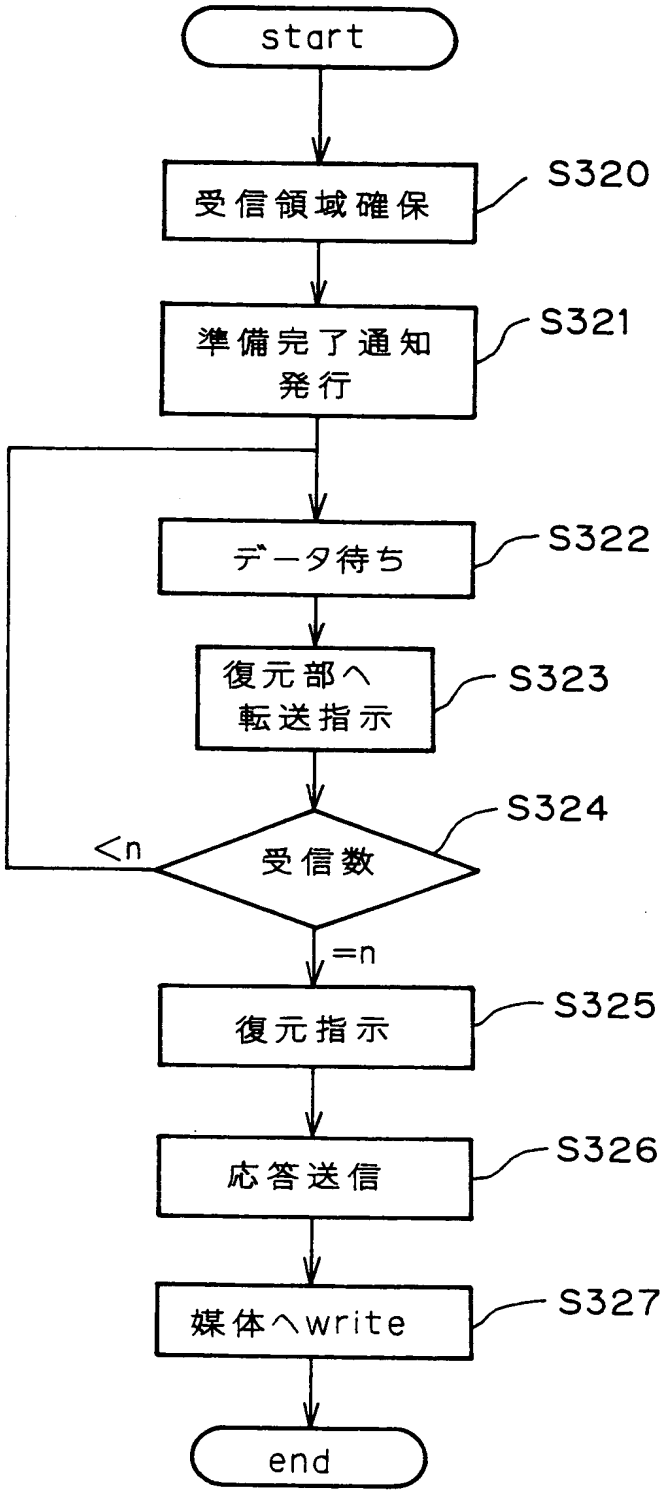
【図１４】


200
201
処理シーケンサ
204
通信部
202
冗長化部
203
復元部
104
IOスケジューラ
106
バッファメモリ
105
媒体制御部
ストレージコントローラ
101
記憶媒体

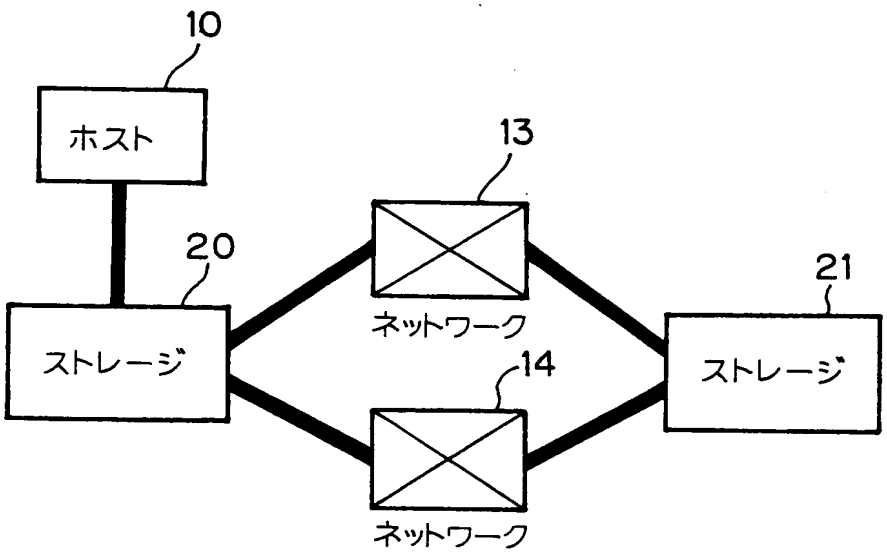
【図１５】


start
転送対象を最初のブロックに設定
S300
冗長化writeコマンド発行指示
S301
媒体read
S302
受信状態待ち read完了待ち
S303
冗長化
S304
データ転送
S305
完了待ち
S306
全転送完了?
S307
Y
N
転送対象を次ブロックに設定
S309
完了通知
S308
end

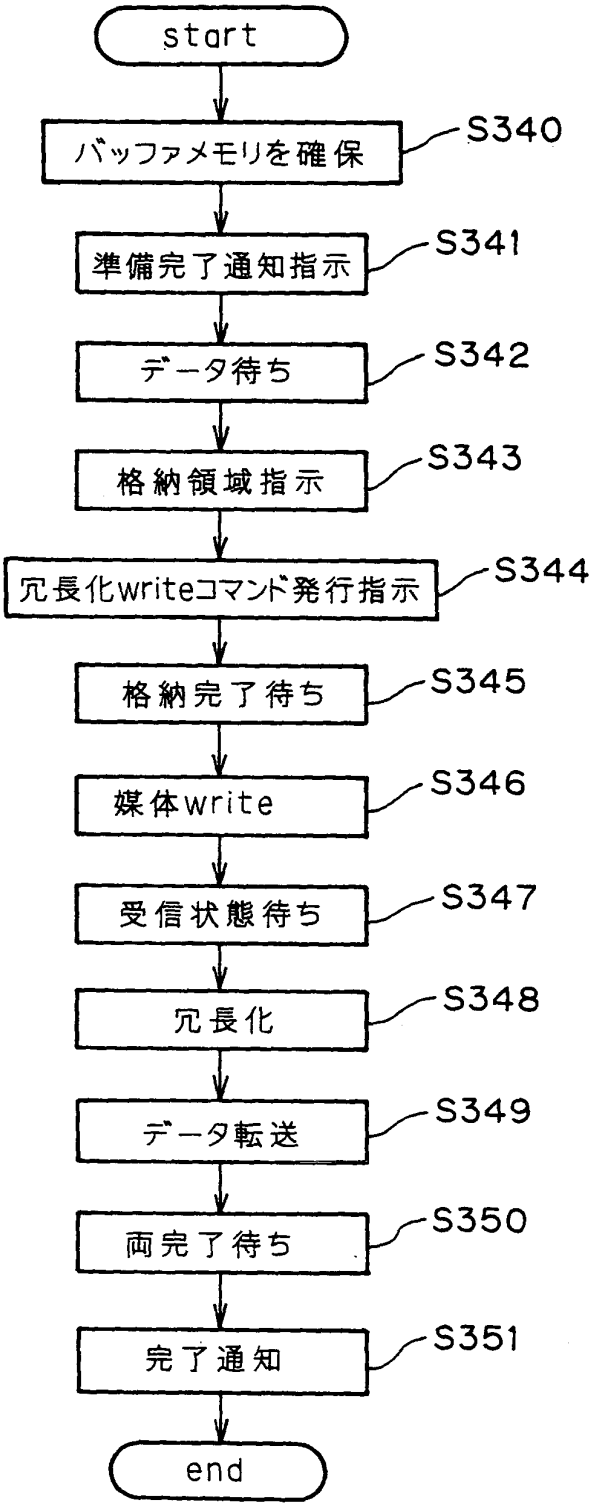
【図１６】


start
受信領域確保
S320
準備完了通知
発行
S321
データ待ち
S322
復元部へ
転送指示
S323
受信数
S324
<n
=n
復元指示
S325
応答送信
S326
媒体へwrite
S327
end

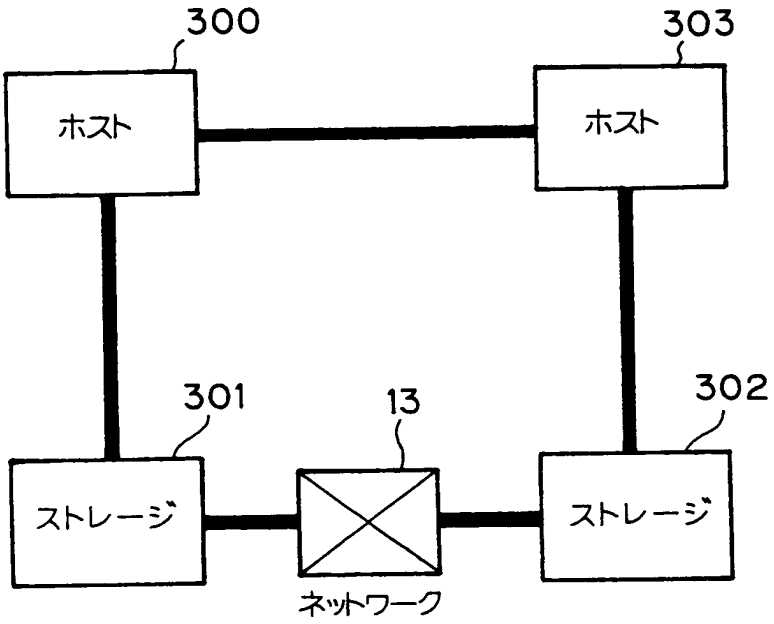
【図１７】



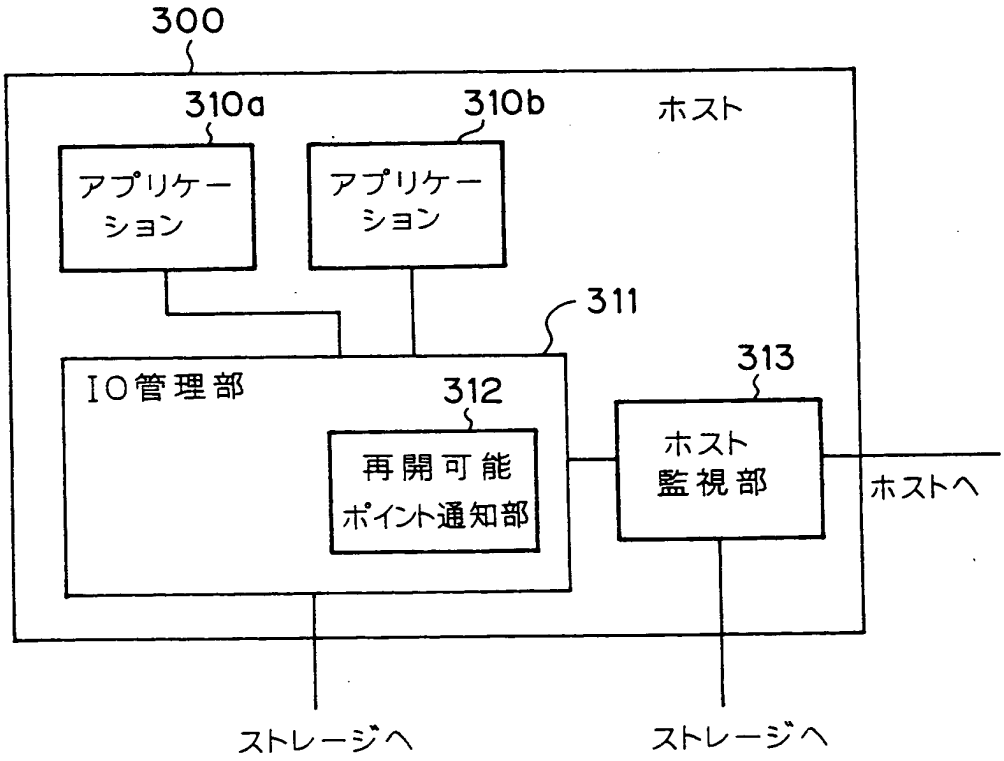
【図１８】


start
バッファメモリを確保
S340
準備完了通知指示
S341
データ待ち
S342
格納領域指示
S343
冗長化writeコマンド発行指示
S344
格納完了待ち
S345
媒体write
S346
受信状態待ち
S347
冗長化
S348
データ転送
S349
両完了待ち
S350
完了通知
S351
end

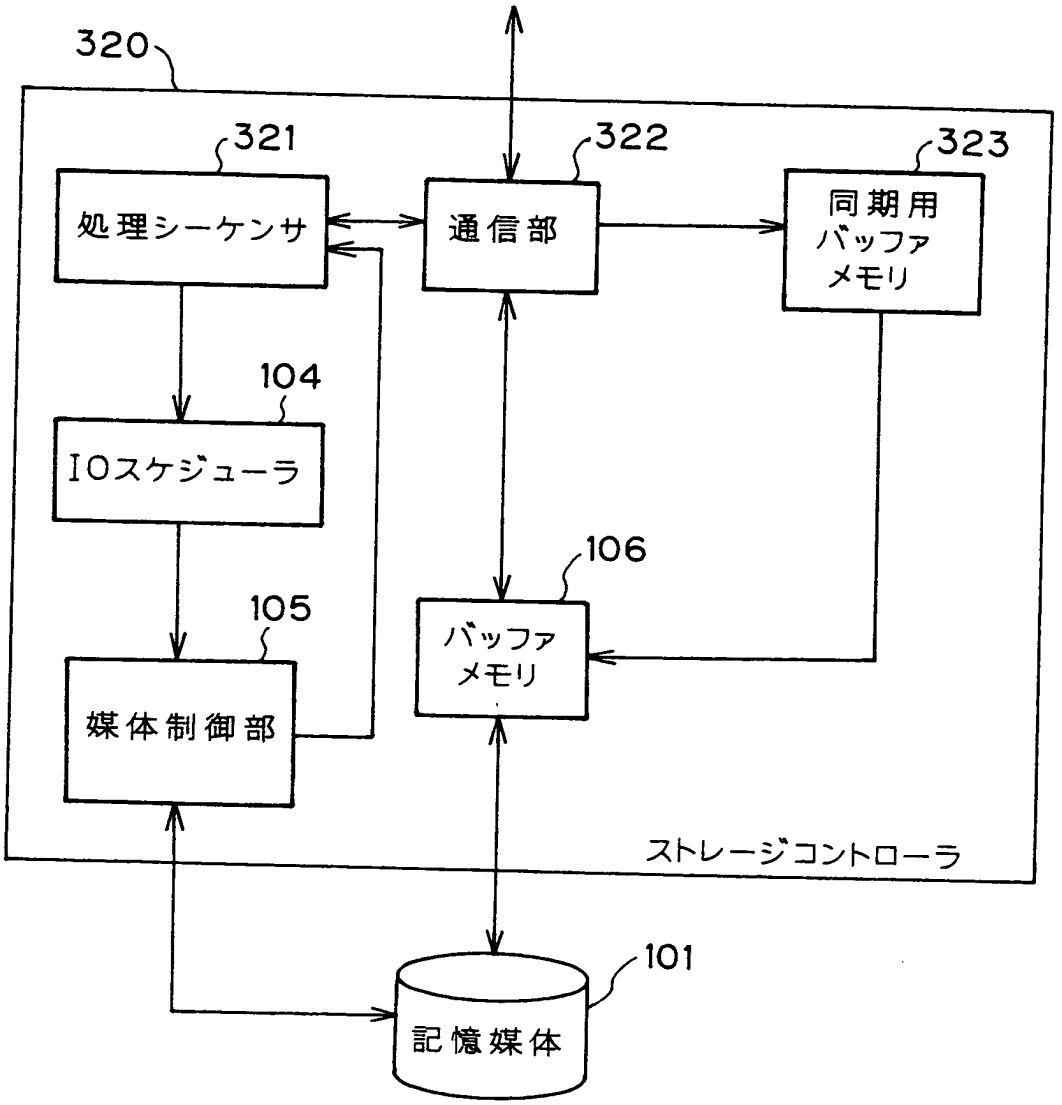
【図１９】


300
303
ホスト
ホスト
301
13
302
ストレージ
ストレージ
ネットワーク

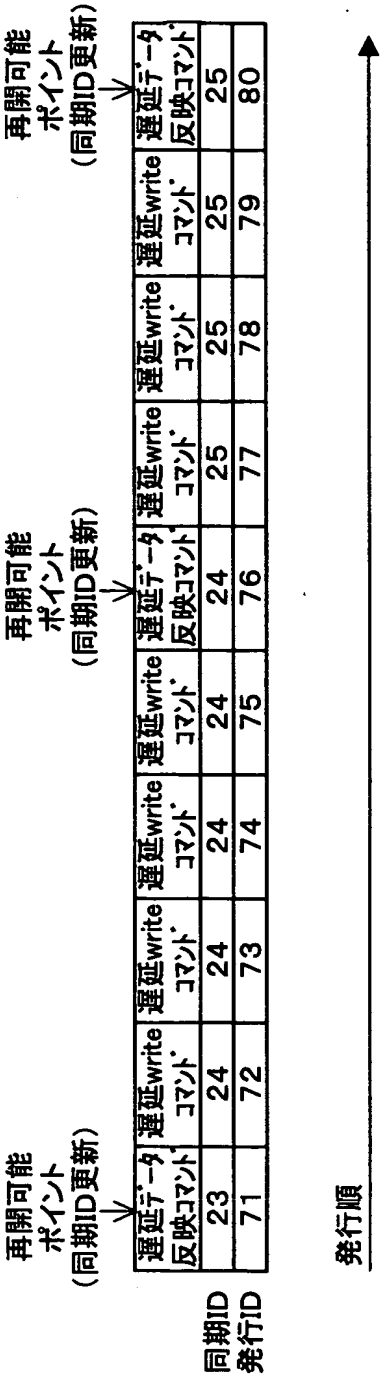
【図２０】


300
310a
310b
ホスト
アプリケーション
アプリケーション
311
313
IO管理部
312
再開可能ポイント通知部
ホスト監視部
ホストへ
ストレージへ
ストレージへ

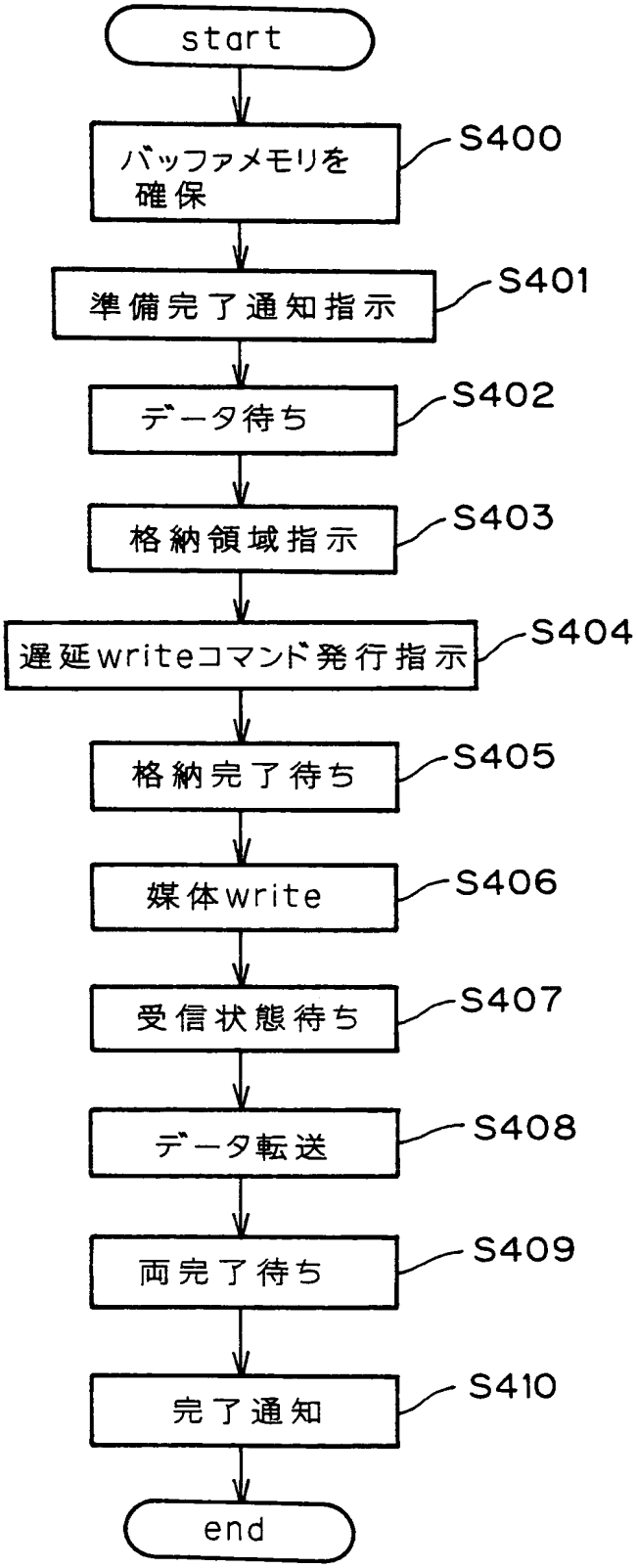
【図２１】


320
321
322
323
処理シーケンサ
通信部
同期用
バッファ
メモリ
104
IOスケジューラ
106
105
バッファ
メモリ
媒体制御部
ストレージコントローラ
101
記憶媒体

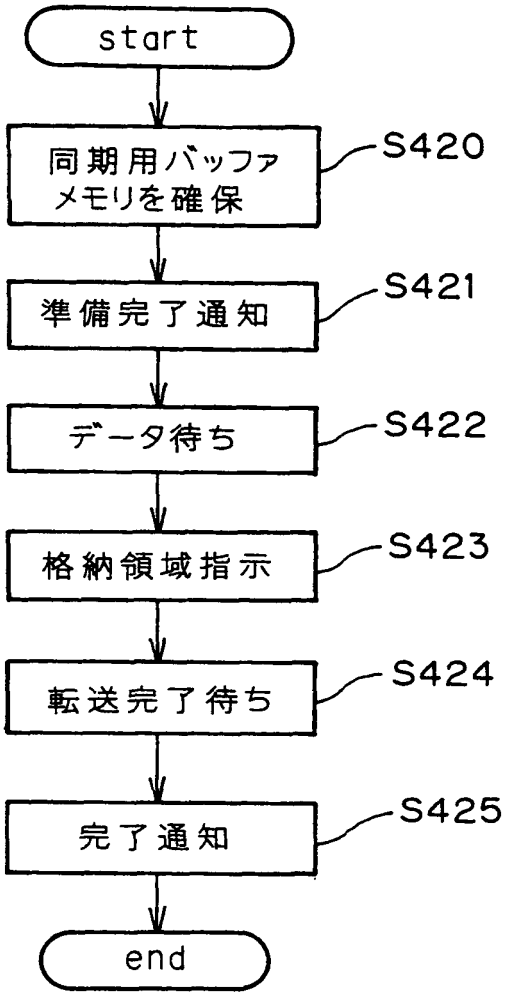
【図２２】


再開可能
ポイント
（同期ID更新）
再開可能
ポイント
（同期ID更新）
再開可能
ポイント
（同期ID更新）
遅延データ
反映コマンド
遅延write
コマンド
遅延write
コマンド
遅延write
コマンド
遅延データ
反映コマンド
遅延write
コマンド
遅延write
コマンド
遅延データ
反映コマンド
同期ID
発行ID
23
71
24
72
24
73
24
74
24
75
24
76
25
77
25
78
25
79
25
80
発行順

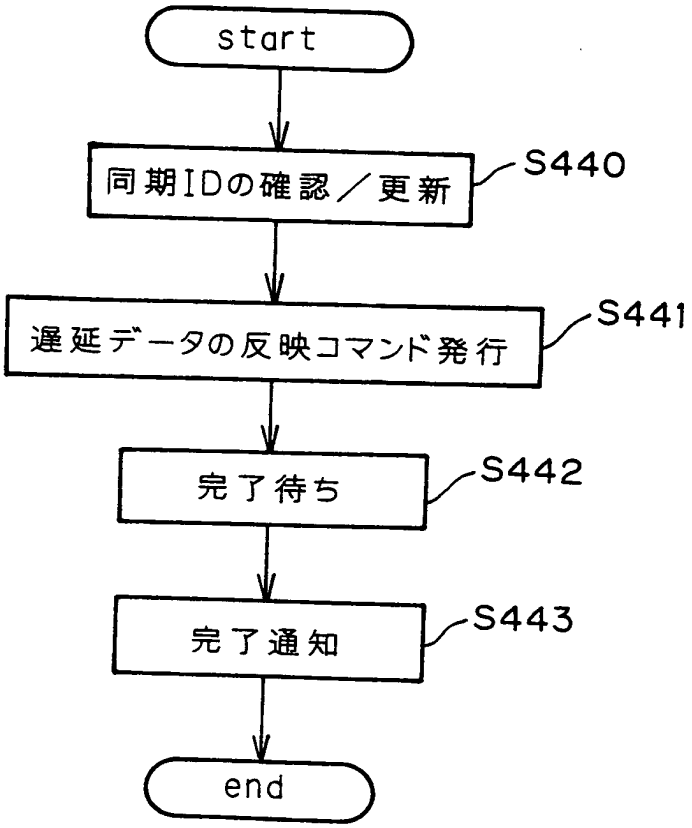
【図２３】


start
バッファメモリを確保
S400
準備完了通知指示
S401
データ待ち
S402
格納領域指示
S403
遅延writeコマンド発行指示
S404
格納完了待ち
S405
媒体write
S406
受信状態待ち
S407
データ転送
S408
両完了待ち
S409
完了通知
S410
end

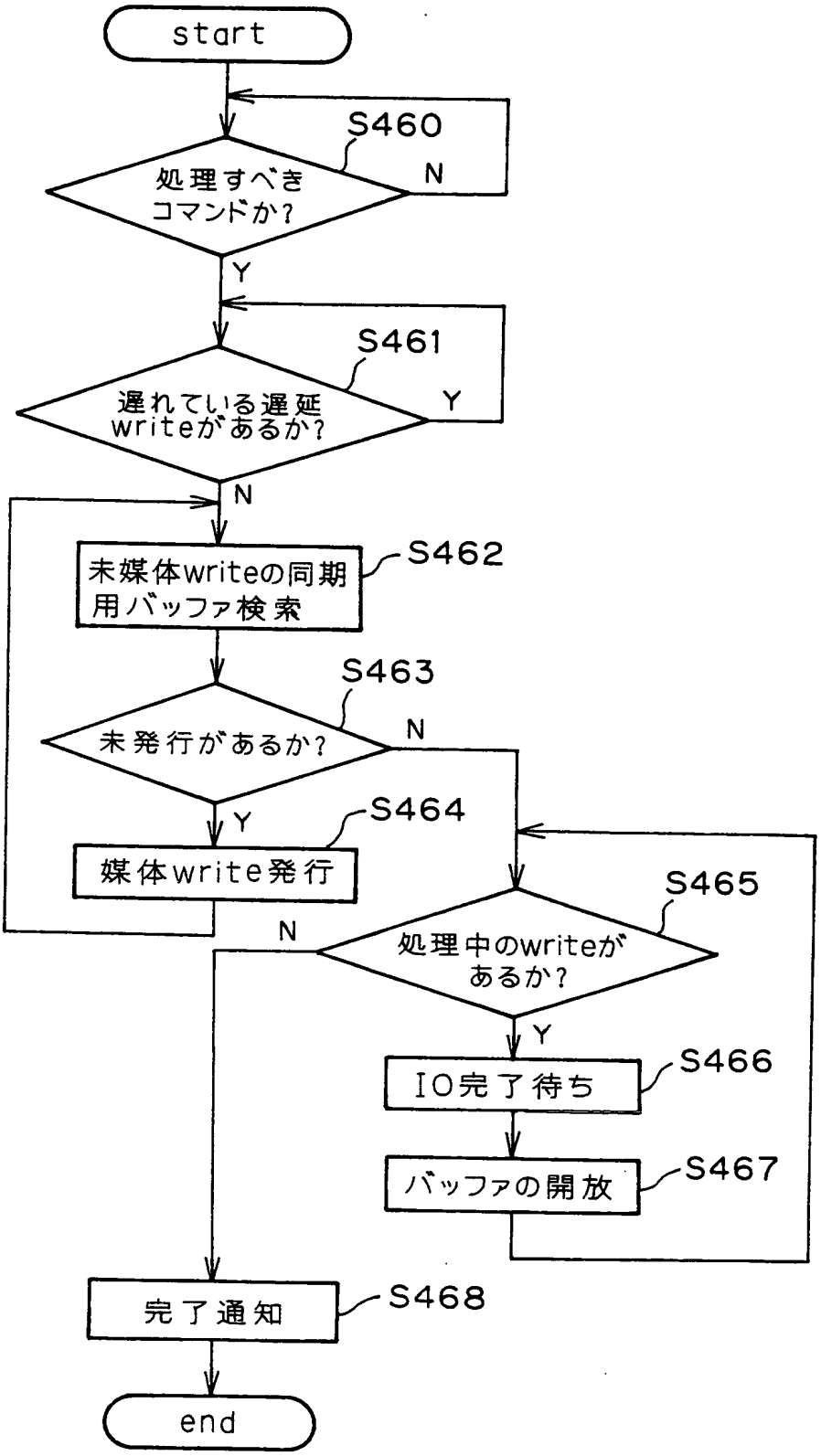
【図２４】


start
同期用バッファメモリを確保
S420
準備完了通知
S421
データ待ち
S422
格納領域指示
S423
転送完了待ち
S424
完了通知
S425
end

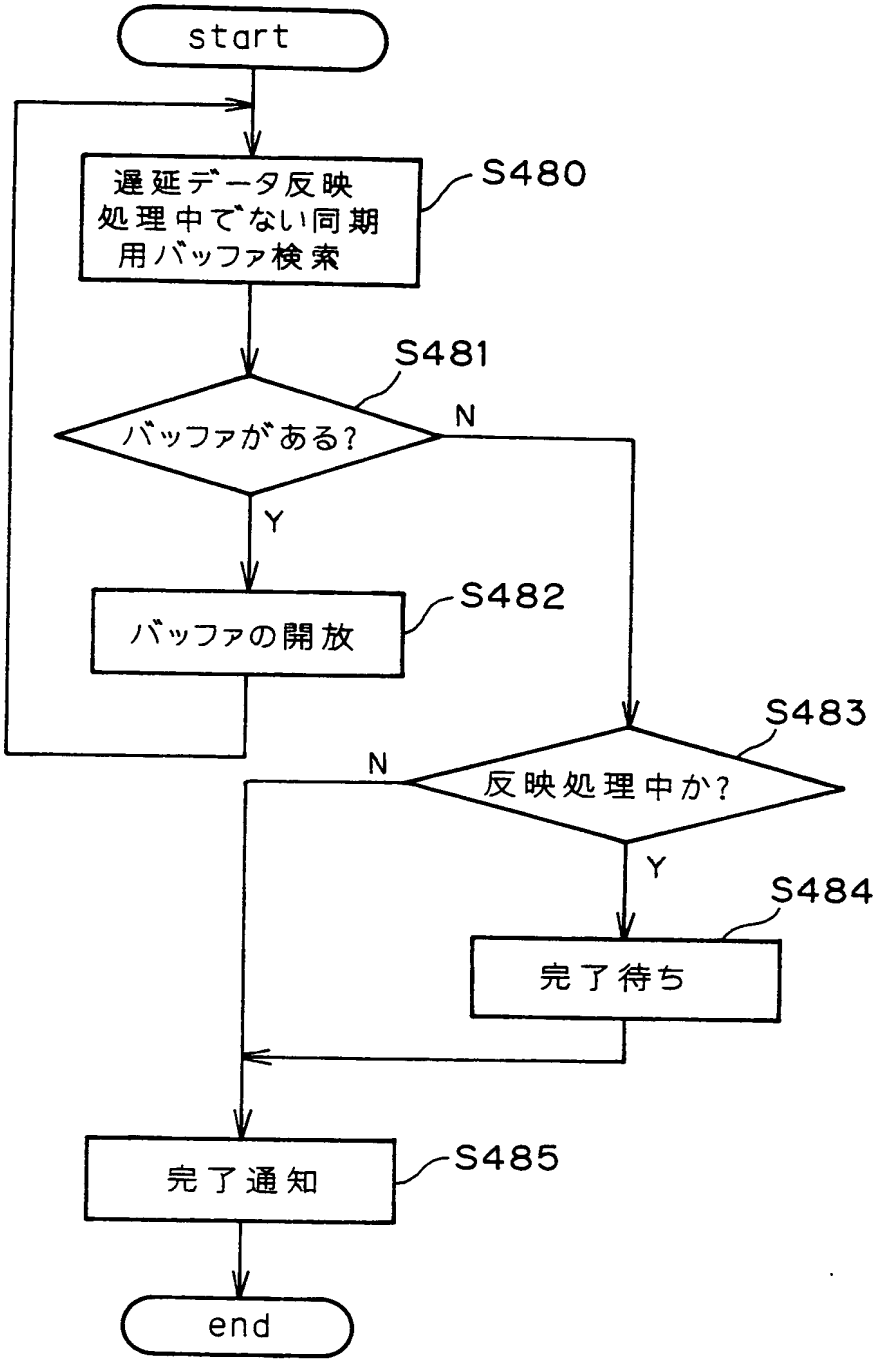
【図２５】


start
同期IDの確認／更新
S440
遅延データの反映コマンド発行
S441
完了待ち
S442
完了通知
S443
end

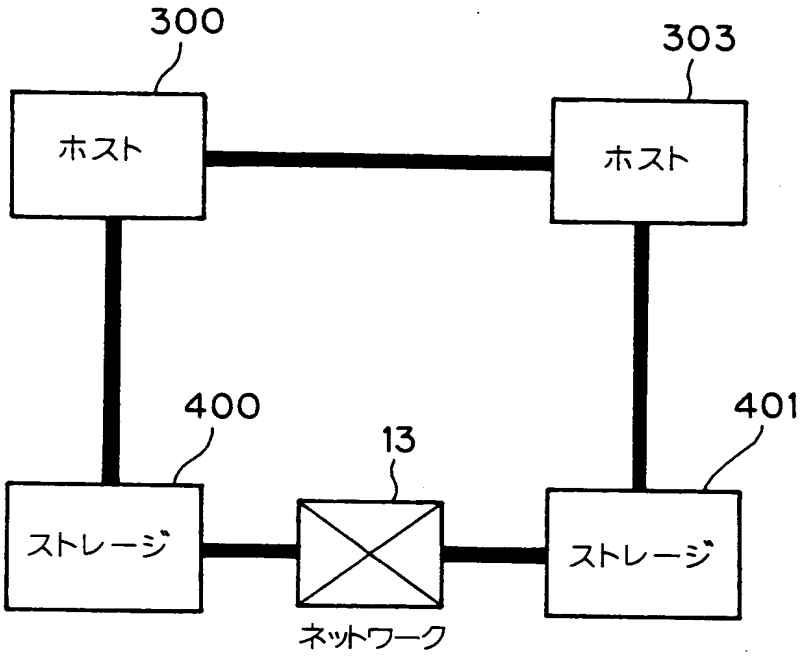
【図２６】


start
S460
処理すべき
コマンドか?
N
Y
S461
遅れている遅延
writeがあるか?
Y
N
未媒体writeの同期
用バッファ検索
S462
S463
未発行があるか?
N
Y
S464
媒体write発行
S465
処理中のwriteが
あるか?
N
Y
IO完了待ち
S466
バッファの開放
S467
完了通知
S468
end

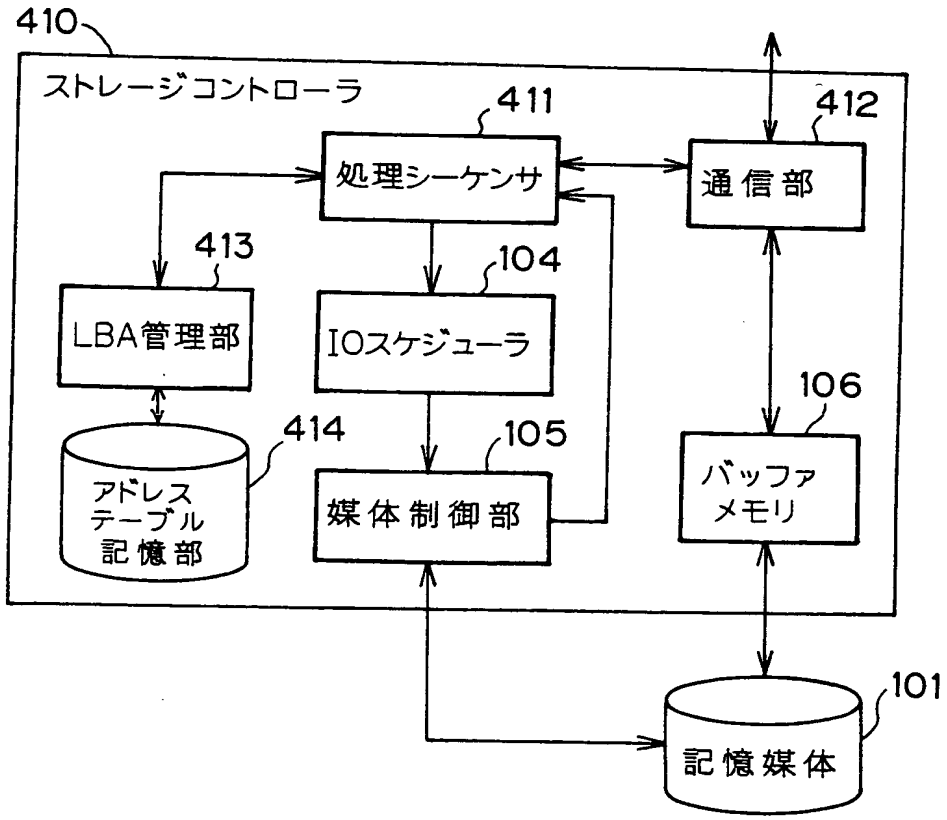
【図２７】


start
遅延データ反映処理中でない同期用バッファ検索
S480
S481
バッファがある?
N
Y
バッファの開放
S482
S483
反映処理中か?
N
Y
S484
完了待ち
完了通知
S485
end

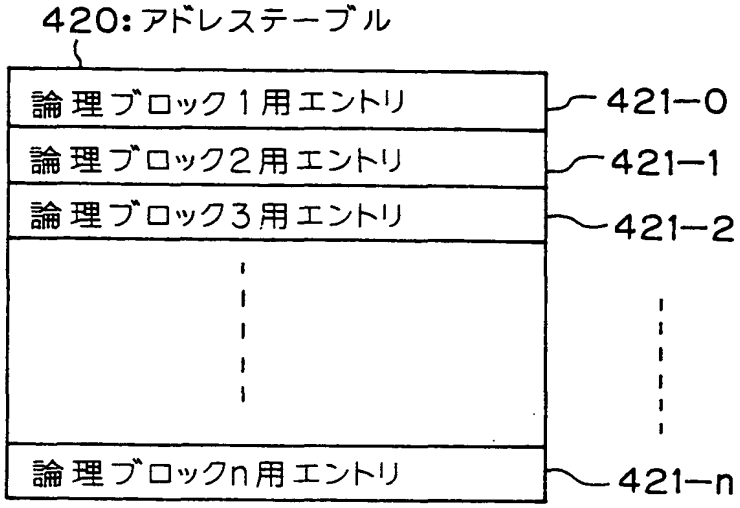
【図２８】


300
303
ホスト
ホスト
400
401
13
ストレージ
ストレージ
ネットワーク

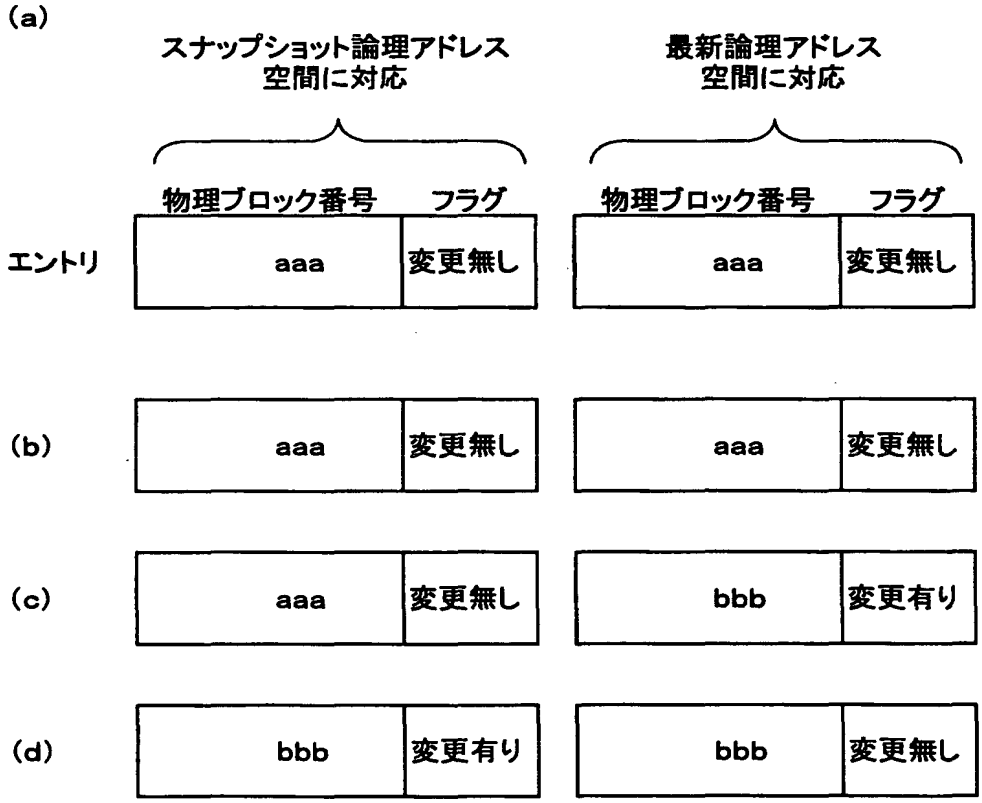
【図２９】


410
ストレージコントローラ
411
412
処理シーケンサ
通信部
413
104
LBA管理部
IOスケジューラ
106
414
105
アドレス
テーブル
記憶部
媒体制御部
バッファ
メモリ
101
記憶媒体

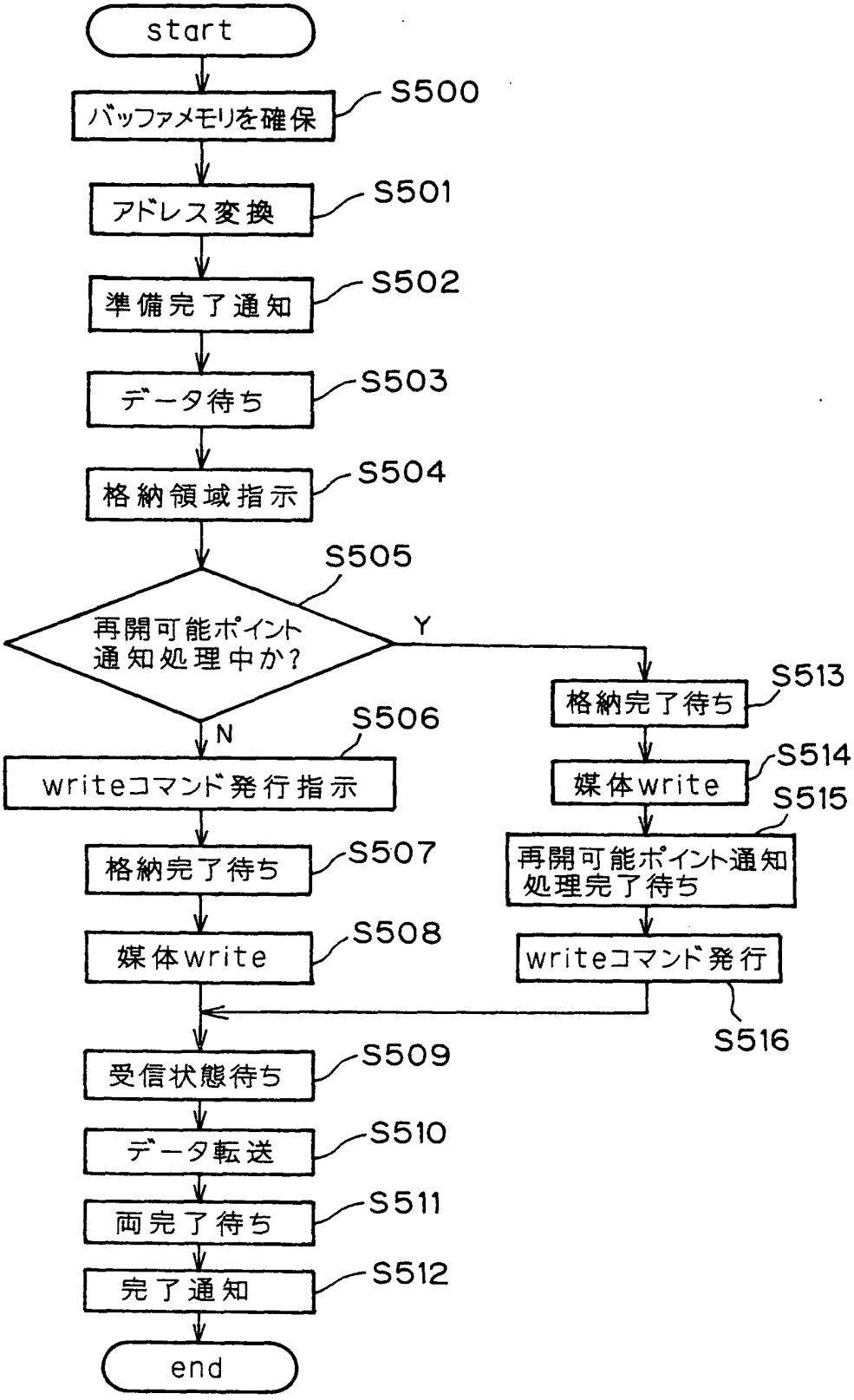
【図３０】


420:アドレステーブル
論理ブロック１用エントリ
論理ブロック2用エントリ
論理ブロック3用エントリ
論理ブロックn用エントリ
421−0
421−1
421−2
421−n

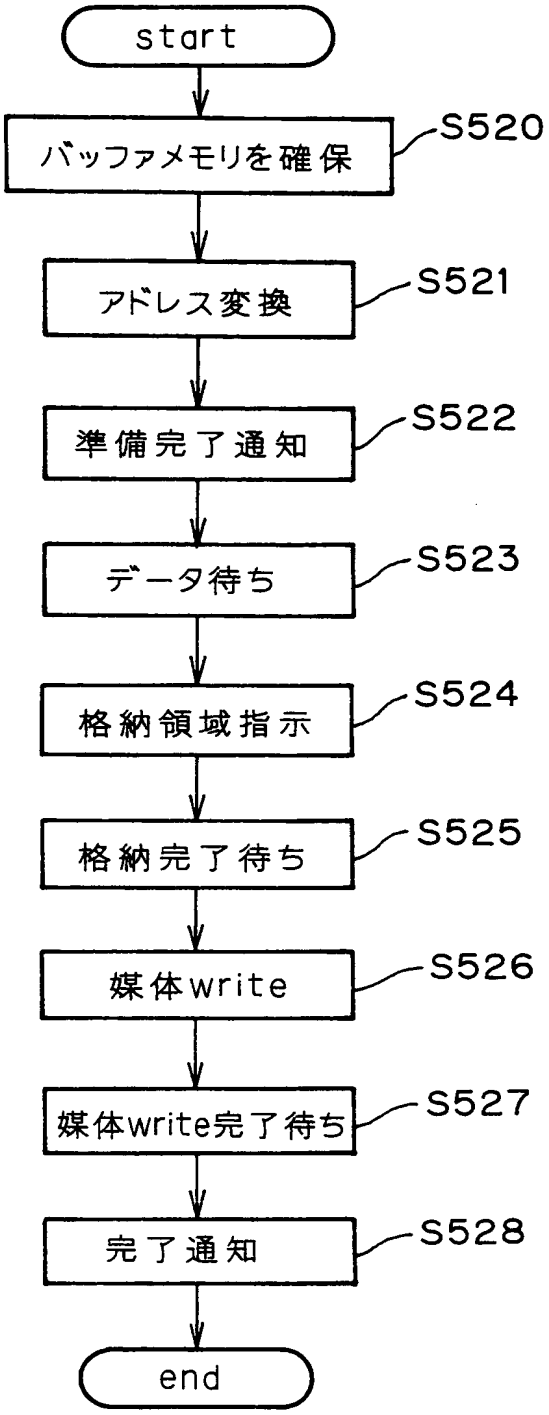
【図３１】


(a)
スナップショット論理アドレス空間に対応
最新論理アドレス空間に対応
物理ブロック番号
フラグ
物理ブロック番号
フラグ
エントリ
aaa
変更無し
aaa
変更無し
(b)
aaa
変更無し
aaa
変更無し
(c)
aaa
変更無し
bbb
変更有り
(d)
bbb
変更有り
bbb
変更無し

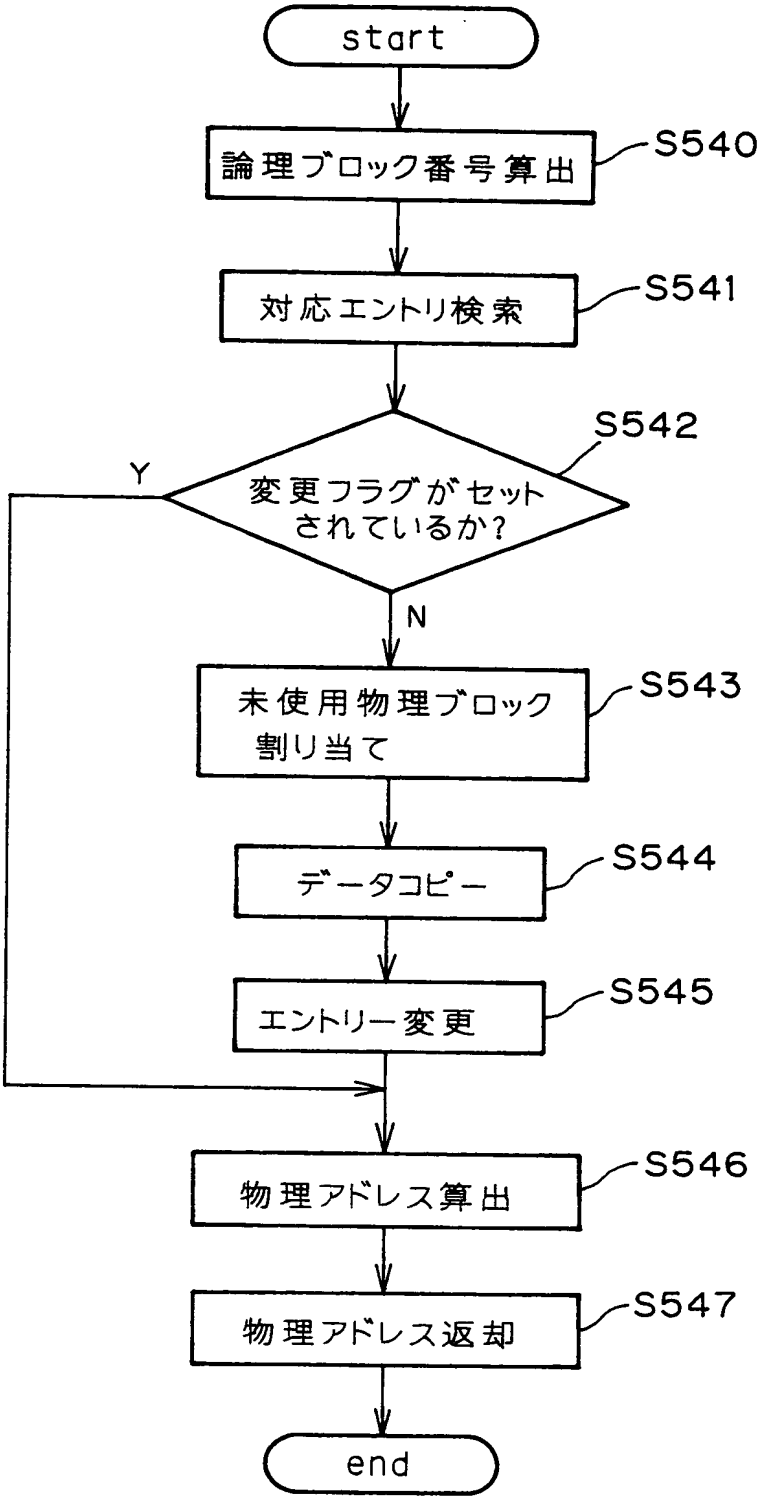
【図３２】


start
バッファメモリを確保
S500
アドレス変換
S501
準備完了通知
S502
データ待ち
S503
格納領域指示
S504
S505
再開可能ポイント通知処理中か?
Y
N
S506
writeコマンド発行指示
格納完了待ち
S507
媒体write
S508
受信状態待ち
S509
データ転送
S510
両完了待ち
S511
完了通知
S512
end
S513
格納完了待ち
S514
媒体write
S515
再開可能ポイント通知処理完了待ち
writeコマンド発行
S516

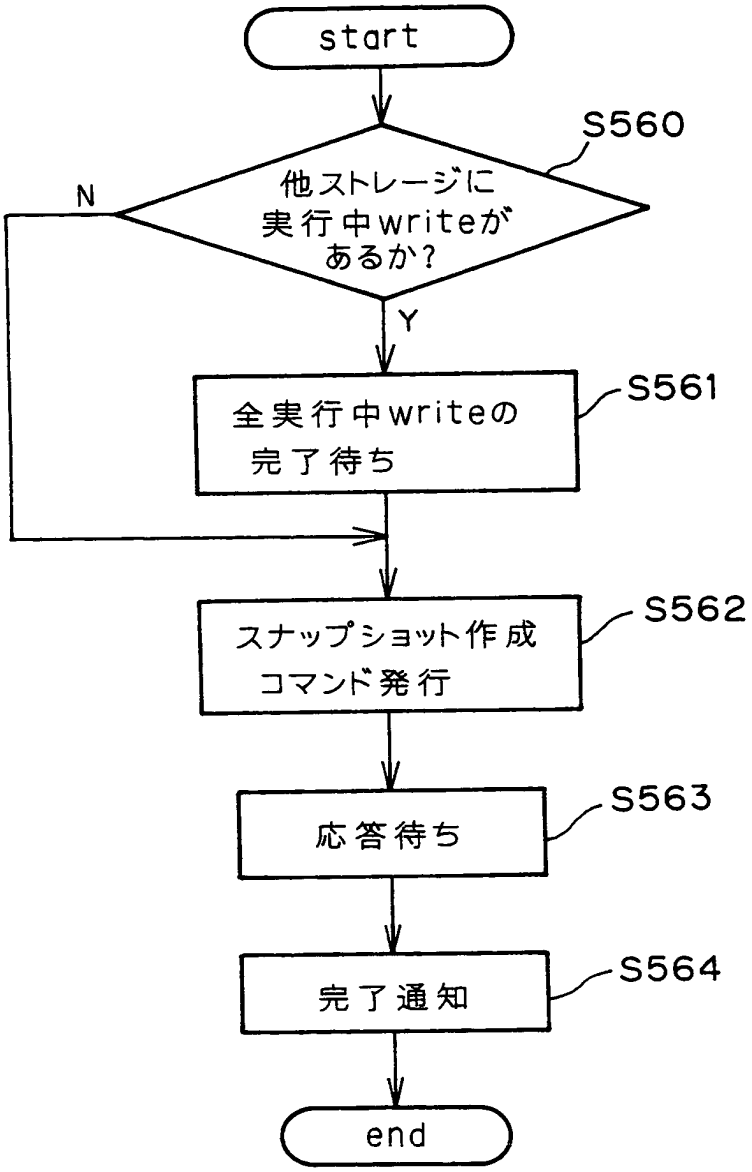
【図３３】


start
バッファメモリを確保
S520
アドレス変換
S521
準備完了通知
S522
データ待ち
S523
格納領域指示
S524
格納完了待ち
S525
媒体write
S526
媒体write完了待ち
S527
完了通知
S528
end

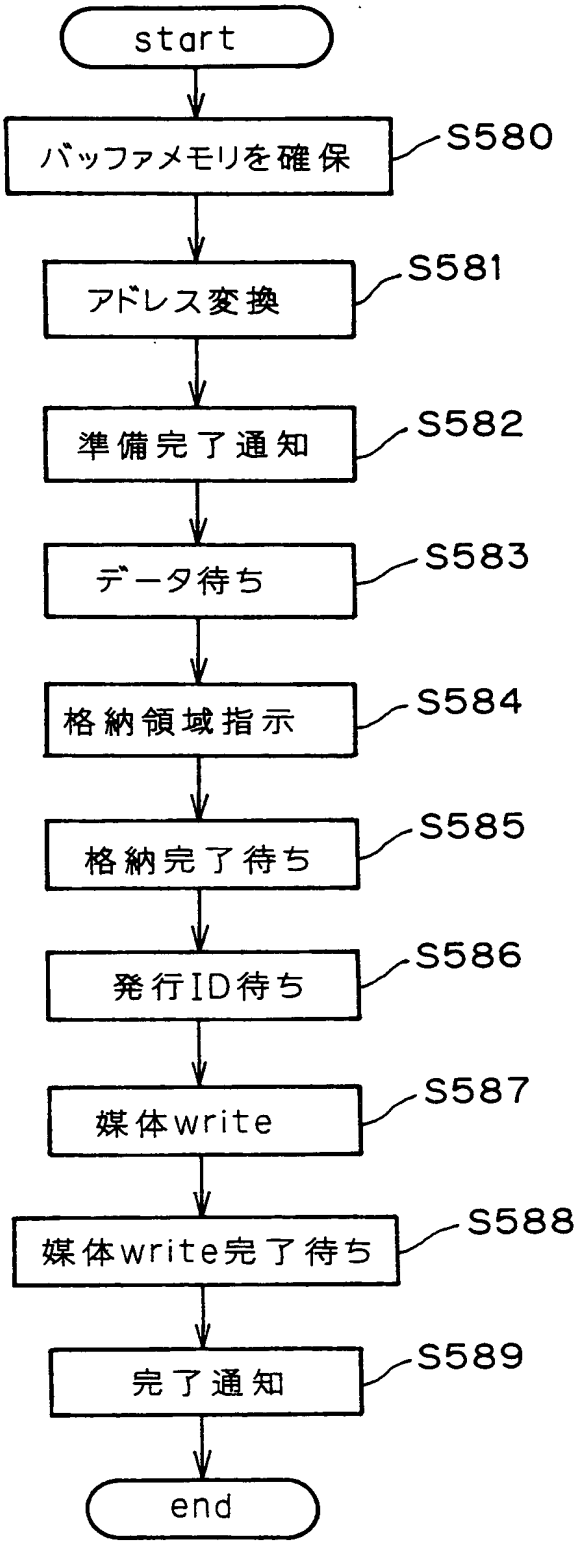
【図３４】


start
論理ブロック番号算出
S540
対応エントリ検索
S541
S542
変更フラグがセット
されているか?
Y
N
未使用物理ブロック
割り当て
S543
データコピー
S544
エントリー変更
S545
物理アドレス算出
S546
物理アドレス返却
S547
end

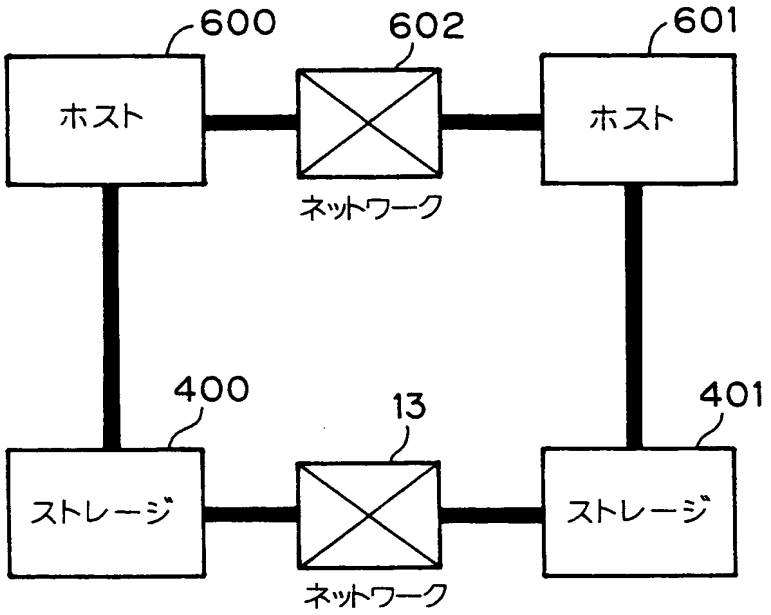
【図３５】


start
S560
他ストレージに
実行中writeが
あるか?
N
Y
全実行中writeの
完了待ち
S561
スナップショット作成
コマンド発行
S562
応答待ち
S563
完了通知
S564
end

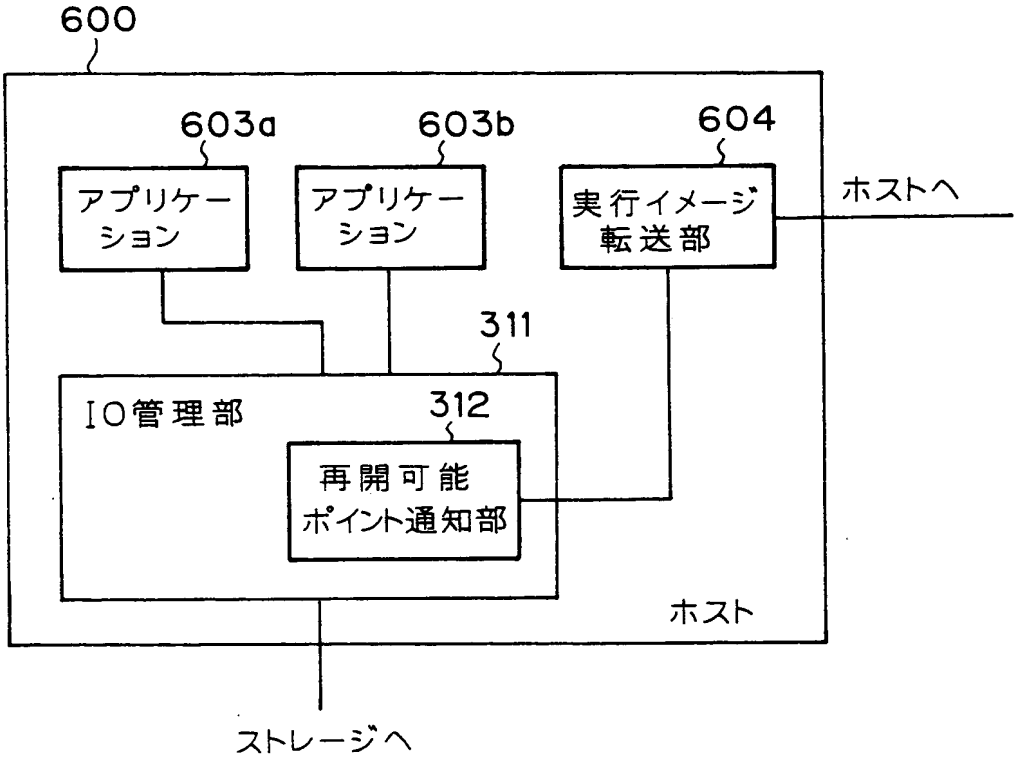
【図３６】


start
バッファメモリを確保
S580
アドレス変換
S581
準備完了通知
S582
データ待ち
S583
格納領域指示
S584
格納完了待ち
S585
発行ID待ち
S586
媒体write
S587
媒体write完了待ち
S588
完了通知
S589
end

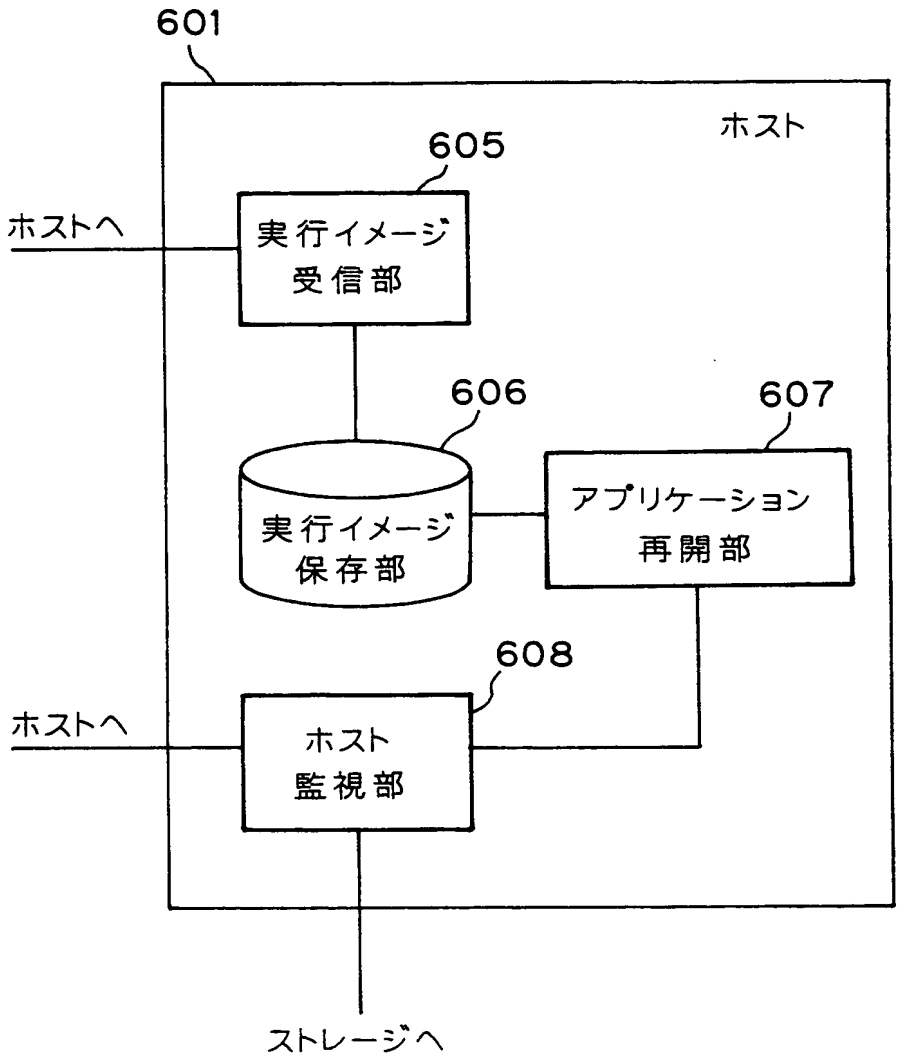
【図３７】


600
602
601
ホスト
ホスト
ネットワーク
400
13
401
ストレージ
ストレージ
ネットワーク

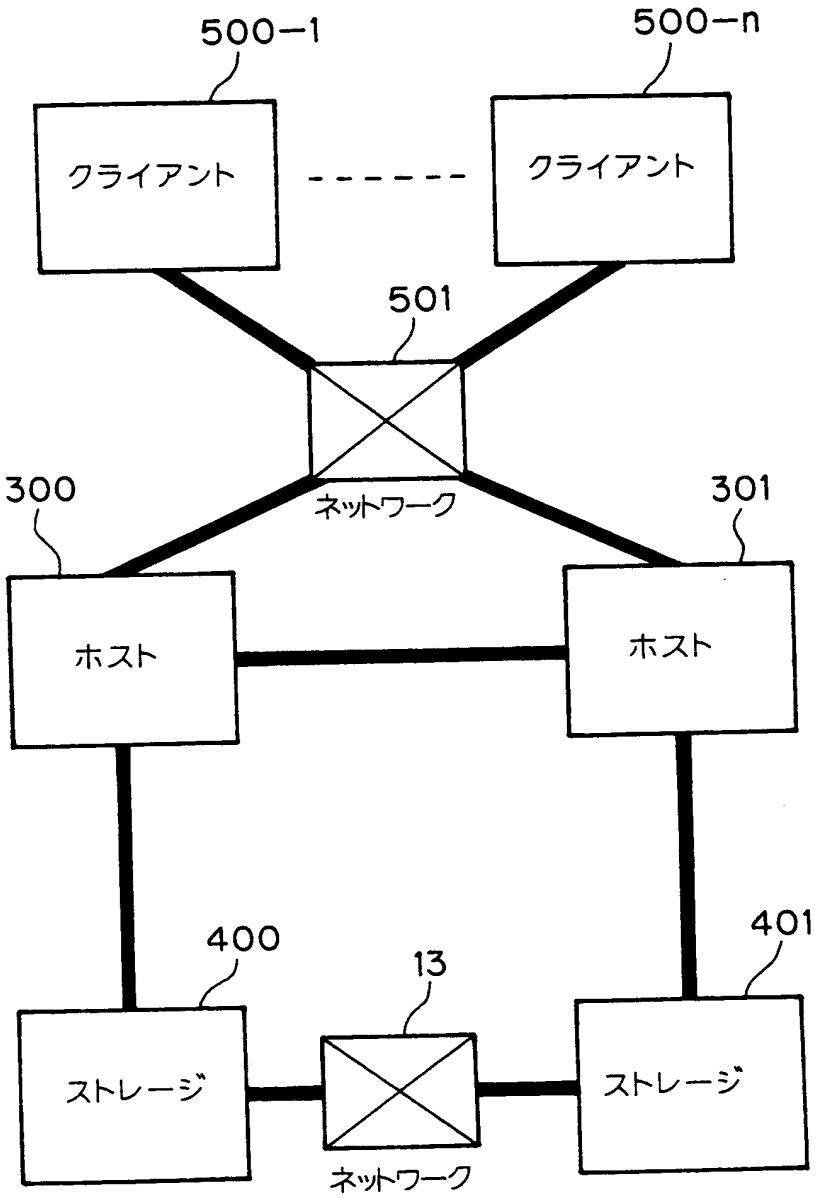
【図３８】


600
603a
603b
604
アプリケーション
アプリケーション
実行イメージ転送部
ホストへ
311
IO管理部
312
再開可能ポイント通知部
ホスト
ストレージへ

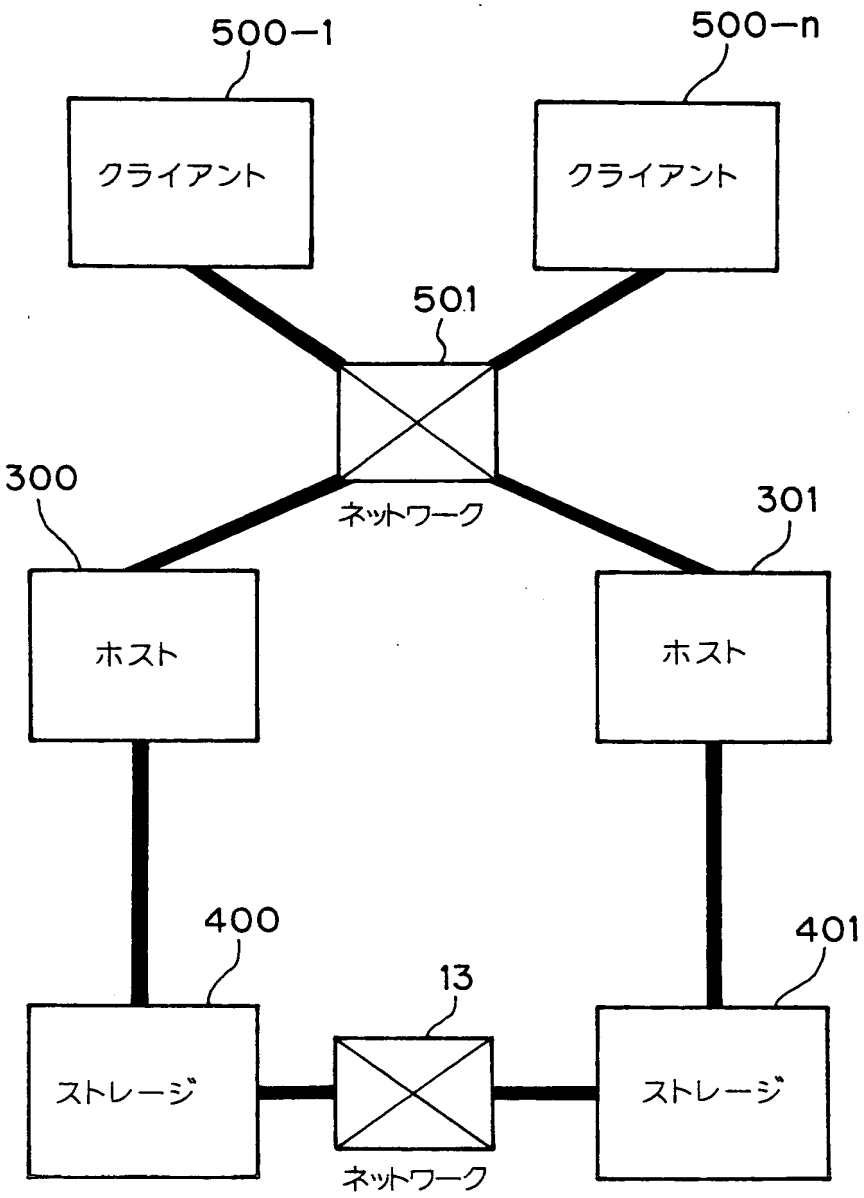
【図３９】


601
ホスト
605
ホストへ
実行イメージ
受信部
606
607
実行イメージ
保存部
アプリケーション
再開部
608
ホストへ
ホスト
監視部
ストレージへ

【図４０】


500−1
500−n
クライアント
クライアント
501
ネットワーク
300
301
ホスト
ホスト
400
13
401
ストレージ
ストレージ
ネットワーク

【図４１】


500−1
500−n
クライアント
クライアント
501
300
301
ネットワーク
ホスト
ホスト
400
401
13
ストレージ
ストレージ
ネットワーク

【書類名】　　要約書

【要約】

【課題】　待機系のシステムにデータを送信して耐障害性を確保する際、データの送信に伴う正常系のシステムの処理の遅れを防止する。

【解決手段】　中継装置１５は、正常系のホスト（ホストコンピュータ）１０およびストレージ１１の設置場所において地震等の災害が発生したときに、災害の影響が波及すると想定される範囲の外に設置される。さらに、ストレージ１１とストレージ１２とが直接データを転送する場合のデータ転送時間よりも、ストレージ１１と中継装置１５との間のデータ転送時間が短くなるような位置に設置される。中継装置１５は、ストレージ１１からデータを受信した場合、ストレージ１２に対してそのデータの送信を完了させる前に、ストレージ１１に対してデータ受信完了を通知する。

【選択図】　図１

出　願　人　履　歴　情　報

識別番号　　〔０００００４２３７〕

１．変更年月日　　１９９０年　８月２９日
［変更理由］　新規登録
住　所　　東京都港区芝五丁目７番１号
氏　名　　日本電気株式会社

出証特２００３－３０３７４２５